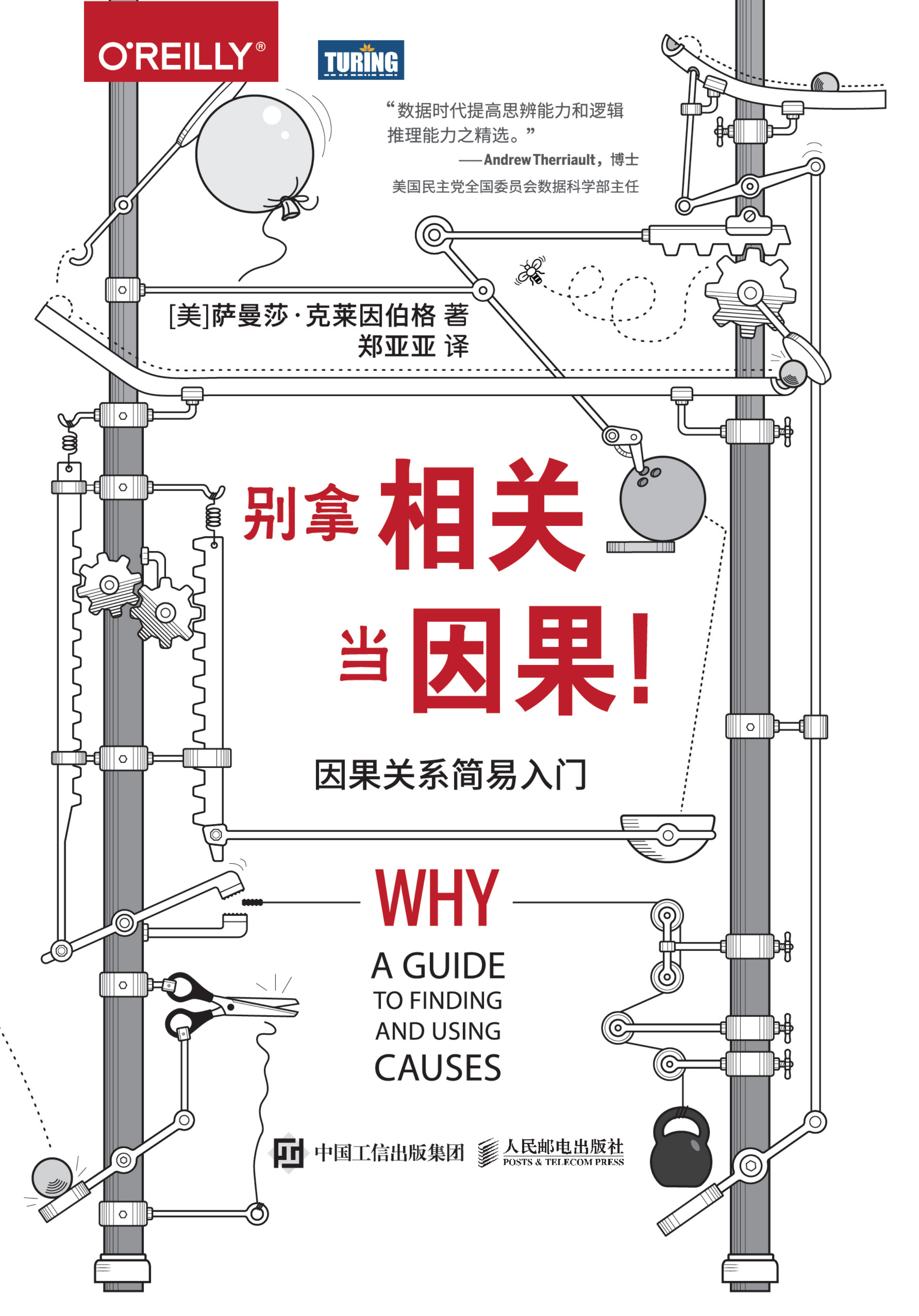
O'REILLY®
TURING



"数据时代提高思辨能力和逻辑推理能力之精选。"

——Andrew Therriault，博士

美国民主党全国委员会数据科学部主任

[美]萨曼莎·克莱因伯格 著
郑亚亚 译

别拿相关当因果！

因果关系简易入门

WHY

A GUIDE TO FINDING AND USING CAUSES

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

译者简介

郑亚亚

上海外国语大学高级翻译学院翻译学专业2016级博士研究生，研究方向为口笔译研究与翻译教学。

封面设计：Anton Khodakovsky 张健

图灵社区：iTuring.cn

热线：(010)51095186转600

数字版权声明

图灵社区的电子书没有采用专有客户端，您可以在任意设备上，用自己喜欢的浏览器和PDF阅读器进行阅读。

但您购买的电子书仅供您个人使用，未经授权，不得进行传播。

我们愿意相信读者具有这样的良知和觉悟，与我们共同保护知识产权。

如果购买者有侵权行为，我们可能对该用户实施包括但不限于关闭该帐号等维权措施，并可能追究法律责任。

别拿相关当因果！因果关系简易入门

Why: A Guide to Finding and Using Causes

[美] 萨曼莎·克莱因伯格　著
郑亚亚　译

Beijing • Boston • Farnham • Sebastopol • Tokyo

O'Reilly Media, Inc. 授权人民邮电出版社出版

人 民 邮 电 出 版 社
北　　京

图书在版编目（CIP）数据

别拿相关当因果！ ：因果关系简易入门 / （美）萨曼莎·克莱因伯格（Samantha Kleinberg）著 ；郑亚亚译. -- 北京 ：人民邮电出版社，2018.7
ISBN 978-7-115-48518-2

Ⅰ. ①别… Ⅱ. ①萨… ②郑… Ⅲ. ①因果性—研究 Ⅳ. ①B025.5

中国版本图书馆CIP数据核字(2018)第108657号

内 容 提 要

本书是写给普通人的因果逻辑入门书，旨在帮助读者培养严谨的思维方式，在不借助任何专业知识的前提下，准确定位问题。主要内容包括：认识原因，对原因的理解和运用，如何只通过观察找到原因，大数据集与原因的关系，因果关系相关实验，如何利用因果关系来制定有效的干预措施，研究因果关系的意义。

本书适合所有对探究事件真相感兴趣的读者，无须统计学等专业背景。

◆ 著　　[美] 萨曼莎·克莱因伯格
译　　郑亚亚
责任编辑　朱　巍
责任印制　周昇亮

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京　　印刷

◆ 开本：880×1230　1/32
印张：9.375
字数：249千字　　2018年7月第1版
印数：1－3 000册　　2018年7月北京第1次印刷

著作权合同登记号　图字：01-2018-0946号

定价：69.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

版 权 声 明

尽管用先进的计算工具很容易从数据中找到规律，但是最深刻的认识还是来自于对这些规律来源的把握，而这可不能只通过计算机来完成。克莱因伯格巧妙地向读者介绍了寻找因果关系过程中所用到的主要概念和方法，思路清晰且内容实用，使得这本书不同于这一领域的其他著作。书中内容全面又易于理解，是科研领域的新人、经验丰富的专家以及其他想要从数据中获取更多认知的读者的必读之物。

——Andrew Therriault，博士，美国民主党
全国委员会数据科学部主任

哲学、经济学、统计学以及逻辑学都有志于理清因果关系，克莱因伯格成功地将这些完全不同的方法以一种简单而又实用的方式综合在了一起。随着我们的生活越来越多地“为数据所驱动”，要想理解国家政策、健康问题以及我们周围的世界，就必须对“从观察中推理因果关系”有一个清晰的思考。

——Chris Wiggins，博士，《纽约时报》首席数据科学家，
哥伦比亚大学副教授

因果关系是生活中的一个重要特征，但是人们对它有着大量争议和误解。本书在未借助任何先验知识或专业技术的前提下，对因果关系做出了清晰的解释，并且通俗易懂、妙趣横生，用严密的逻辑和深刻的分析来帮助我们理解复杂的概念。

——David Lagnado，博士，伦敦大学学院高级讲师

目　　录

前　　言

喝咖啡会使人长寿吗？是谁把流感传染给你的？股票价格为什么会上涨？无论是做饮食安排还是投资选择，抑或是责怪某人毁了你的周末，你都需要不断去了解其中的原因。正是这种因果关系在帮助我们预测未来，解释过去，让我们能够介入其中并对事物的变化产生影响。与流感病人接触会让你在一段时间内也染上流感，知道了这一事实，你就能知道自己会在什么时候出现流感症状。针对性很强的游说可以为你筹集政治竞选资金，了解了这一关系，你就可以将这些游说活动视为增加竞选资金的一个可行方案。高强度运动会导致高血糖，意识到这一点，你就可以帮助糖尿病患者控制血糖。

尽管推断因果关系这一技能非常重要，但你之前可能没有接触过这方面的课程。事实上，你可能都未曾静下心来想过，为什么某件事会成为另一件事发生的原因。虽然这个问题的答案涉及很多因素，但从根本上看，原因可以提高一个事件发生的概率，是产生某种结果的前提，或者是让某件事情发生的策略。但是，不能因为某种药物会引发心脏病，就认为某个人的心脏病发作是由这种药物引起的。也不能因为某个地区在缩小班级规模以后，学生的成绩都得到了提高，就认为同样的做法在其他地区也会产生同样的效果。本书不仅要讨论在一切进展顺利的情况下可能会出现哪些

结果，还要研究为什么成功看似很难被复制。除此之外，我们还要考察那些在理论研究中经常被忽视的实际问题。

研究因果关系的方法有很多种（有些是互补的，有些是对立的），而且涉及众多领域（包括哲学、计算机科学、心理学、经济学、医学等）。我无意在这些争辩中选择立场，只想为读者呈现各种观点，厘清各种观点之间的共识与分歧。除此之外，我们还将探讨关于因果关系的心理学（人们是如何了解原因的）、如何进行因果关系的实验（以及这些实验的局限性是什么），以及如何根据因果关系来制定相应的策略（我们是否应该减少食物的含盐量，以此来预防高血压）。

我们首先要弄清楚什么是原因，以及为什么我们常常会弄错事情发生的原因（第 1~3 章）。然后，要认识到在原因的理解和运用方面，“什么时候”与“为什么”同等重要（第 4 章）。接着，要学习如何只通过观察就找到事情发生的原因（第 5 章）。

大型数据集可以让我们找到事情发生的原因，而不是简单用来检验我们的假设。但是我们必须认识到，并不是所有的数据都能用来推理事件发生的原因。在第 6 章，我们将考察数据特征对推理的影响。第 7 章将探索在可以做实验的情况下，如何去克服这些数据特征给我们带来的挑战。这里所说的实验可能是复杂的临床试验，也可能只是某人对自己不同锻炼计划所做的对比实验。通常情况与个别情况之间的差异，正是我们需要使用专业性策略对各种事件做出解释的原因（第 8 章讨论的内容）。但是，要想利用因果关系来制定有效的干预措施，如在菜单上提供食品热量信息来降低肥胖症发生的概率，就需要有更多的信息，而且很多干预措施还可能带来意想不到的后果（第 9 章将详述这一点）。本书将会告诉你为什么因果关系如此难找（比报纸文章告诉你的要更细致、更复杂），以及为什么尽管如此，它仍是一个相当重要且广泛适用的话题。

虽然困难重重，但也并非毫无希望。你将会形成一套基于原因的思

考体系：要问的问题、应引起怀疑的危险信号以及证实因果关系的方法。除了帮你找到事情发生的原因以外，本书还能帮你基于因果关系来做出决策、制定策略，并通过进一步测试来验证你找到的原因。

这本书是为普通读者而写的，我并未假定这些读者具有相关的背景知识。我唯一假定的是读者对因果关系充满好奇，我要让复杂的因果关系变得通俗易懂、广为人知。读完之后，我们会更加关注人们的直觉以及如何从概念上理解因果关系，而不是数学细节（实际上，本书不会介绍任何数学知识）。如果你是计算机科学或统计学博士，也许会学到一些新的工具并且很享受在其他领域的工作之旅，也可能会向往更多方法论方面的知识。不过，本书要研究的只是普通人眼中的因果关系。

第 1 章　引子

因果关系的概念以及寻找因果关系的方法从何而来？

1999 年，一个名叫 Sally Clark 的英国律师被法庭判定谋杀了她的两个孩子。1996 年 12 月，她的第一个儿子在 11 周大的时候突然死亡。当时，人们认为孩子是自然死亡。但是就在第一个孩子夭折一年多以后，Clark 的第二个儿子又在 8 周大的时候死亡了。在这两个案件中，两个孩子似乎都没有什么生理上的疾病。于是，他们的突然死亡引起了人们的怀疑。

这两个案件有很多共同之处：孩子们死的时候年龄差不多，他们的死都是由 Clark 发现的，当时家里只有 Clark 和孩子在一起，而且验尸报告表明两个孩子身上都有伤。一开始，人们认为第一个孩子的伤是抢救时造成的。但是，第二个孩子死后，人们对孩子的伤重新做了检查，而这一次他们认为这些伤很可疑。第二个孩子死亡四周后，警方逮捕了孩子的父母，随后 Clark 被指控为谋杀罪并获刑。

同一个家庭的两个婴儿都死于婴儿猝死综合征（SIDS）的概率有多大？据英国检察官称，发生这种情况的可能性微乎其微，所以这两起死亡一定是谋杀所致。这一论据（一个原因的可能性几乎为零，所以一定还有另外一个原因）造成了这桩著名的冤案。这也是因为统计不当和忽视因果关系而造成严重后果的一个重要案例。

统计学家和因果关系研究者都知道这个案例，其主要原因在于，检

方的论据本质上基于这样的逻辑：被告的辩词几乎不可能为真，所以一定是假的。检方为此请来了一位专家证人——Roy Meadow 博士。Meadow 称，同一个家庭发生两起 SIDS 事件的概率为七千三百万分之一。检方由此认为，因为这个概率非常低，所以这两起死亡事件不可能出于自然原因，一定是谋杀所致。

然而，这一统计数据完全是错误的。即使这个数据是正确的，也不应该这样用。有一份研究报告估算出发生 SIDS 的概率为 1/8543。Meadow 根据这个研究报告提出：同一个家庭发生两起 SIDS 事件的概率为 $1/(8543 \times 8543)$，即约七千三百万分之一。[1] 这种计算方法错误的原因在于，它假定这些事件是相互独立的。抛硬币的时候，无论硬币落地时是正面朝上还是反面朝上，都不会对下一次结果产生任何影响。因为每一次硬币正面朝上的概率都是 1/2，所以将第一次正面朝上的概率与第二次正面朝上的概率相乘所得出的结果就是连续两次正面朝上的概率，这从数学角度来讲是没有问题的。Meadow 当时也是这样计算的。

引发 SIDS 的原因还不确定，但一个很重要的影响因素是孩子所处的环境（比如家里是否有人吸烟和饮酒）。这意味着，如果一个家庭发生过一起 SIDS，那么这个家庭发生第二起 SIDS 的概率就会远大于 1/8543，因为这些孩子的生活环境和遗传基因都是相同的。也就是说，第一起死亡事件会向我们透露第二起死亡事件发生的概率。这个案例和一名演员获得两次奥斯卡金像奖的情况非常相似。金像奖并不是随机颁发的，演员第一次得奖时具备的品质（才华、知名度、人脉）会提高他再次得奖的可能性。

这就是 Clark 一案的症结所在。在这个案子中，两起事件并不是相互独立的，可能还是共同的原因引发的。因此，不能通过简单的乘法来计算这两起事件发生的概率。相反，在计算第二起事件发生的概率时，应该考虑到第一起事件的发生。所以我们需要知道的是，在一个已经发生过一起 SIDS 事件的家庭中，发生第二起 SIDS 事件的概率。本案中的概率在计

算和使用方面都存在十分严重的问题，为此，被告在第一次上诉时请来了一位统计学家作为专家证人，皇家统计学会还专门写了一封信表达了他们对这个案件的关心。[2]

不过此案的问题并不仅仅是误算概率那么简单。在整个案件中，检方试图将事件（即这两起 SIDS 死亡事件）发生的这七千三百万分之一的概率等同于 Clark 无罪的概率。这种错误的推理将事件发生的概率当成了被告有罪或者无罪的概率，这就是我们所说的检察官谬误。[3]

我们知道，一个几乎不可能发生的事件真的发生了。一个家庭中发生两起 SIDS 死亡事件的可能性很小，但是一个家庭中两个婴儿都夭折的可能性也很小。人们不单单会考虑 SIDS 这一解释是否合理，更重要的是，他们会将其与关于这个事件的其他解释进行比较。因此在这个案件中，最好将同一个家庭中两个孩子都被谋杀（检方的假设）的可能性与同一家庭中两个孩子都患了 SIDS 的可能性进行比较。

一个家庭中两个孩子都死于 SIDS 的概率与这两个孩子都感染的概率是不同的。关于这个案件我们还有其他的证据，比如物证和犯罪动机等。必须将这些证据与概率结合起来看（比如说，一个人如果没有犯罪动机、没有作案机会或者没有行凶武器，那么他杀人的概率肯定要低于总谋杀率）。[4]

最后，无论一件事情发生的概率有多低，只要尝试的次数足够多，最后一定会发生。Clark 一案中，那个误算出来的极低的概率（七千三百万分之一）比中百万大博彩的概率（二亿五千八百万分之一）还要高三倍多。一个人中大奖的概率是极低的，但是如果我们说某个地方的某个人会中大奖，这个概率又如何呢？那就高得多了。这就说明，仅通过概率来判断一个人的清白一定会导致一些冤案。这是因为虽然对某个特定的家庭来说，发生这种事件的可能性很小，但是世界上有两个孩子的家庭有上百万个，这种事件总会在某个地方的某个家庭发生。

2003 年 1 月，Clark 第二次上诉时终于翻案。然而，那时她已经在监狱服刑三年了。

为什么 Clark 案会成为因果推理失败的重要案例呢？尽管此案在计算概率的过程中存在很多问题，但最根本的原因是，此案试图用一个事件发生的概率来支撑某个特定的因果结论。“这只是巧合而已”“这个概率有多大”，当你在说服别人相信某个因果关系时，是否也说过这样的话呢？生活中经常有这样的推理：公司来了一名新员工，而同一天你的订书机不见了；一名巫师知道你最喜欢的女性亲属的名字以“M”开头；两名重要人证记得那名嫌犯穿的是一件红色法兰绒衬衫。但是，如果因某件事情不大可能发生，而说其唯一合理的解释就是因果关系，那一定是错误的。前面已经说过，一个不大可能发生的事件在某个人身上发生的概率也许极低，但是在某个地方发生的概率却不低。除了会造成冤案以外，错误的因果推理还可能会带来其他严重的后果，比如将大量的时间和精力浪费在绝不可能起作用的药品上，或者制定一些无用的、代价高昂的公共政策。

本书的目的是提高读者的因果推理能力。严谨的因果思维是指质疑假设、衡量证据、分析各种说辞，以及辨别我们无法得知事情发生原因的情况。有时我们可能无法获得足够的信息来建立因果联系，有时我们获得的信息可能并不是我们所需要的，但重要的是能够认识到这些问题，并与其他人就这些问题进行交流。通过阅读本书，我希望读者至少能够对他们所听到的各种因果推论多一些质疑（我们将讨论在因果推论中需要注意哪些危险信号，以及可以提出哪些问题来衡量这些推论），但首先我们会教大家如何寻找事件发生的原因、如何为因果关系提供强有力的证据，以及如何使用因果关系来指导我们日后的行为。

1.1　何为原因

试着花点时间，给“原因”下个定义。

如果你与上我的因果关系推理课的学生一样，那很可能定义下到一半就开始用各种可能的异议打断自己了。也许你用了“绝大多数时候”或“但并不总是这样”或“只有……”这样的字眼来限定自己的定义。而且你的定义很可能包括一些特征，比如：原因会**导致**某种结果、会使某种结果**更有可能出现**、具有产生某种结果的**能力**，或者会**形成某种结果**。这些特征体现了人们的一种普遍想法：事情的发生都是有原因的，否则它就不会发生。

尽管这种想法并不适用于所有情况，但在本书中，“原因”一词一般是指：它使某种结果更有可能出现，并且没有它某种结果就不会出现或者无法出现，或者说它能够在适当的环境下产生某种结果。

“原因”最早的一种定义来自亚里士多德，他认为原因是用来回答“为什么”的。[5] 所以，如果我们问为什么某件事是这样的，人们可能会解释这个现象是如何产生的（比如水加热后会产生蒸汽）、这个事物的成分是什么（比如氢气和氧气结合会形成水）、这个事物是什么样的（比如椅子的本质就是高出地面的、有靠背的、用来让人坐的东西），或者为什么要做这件事（比如疫苗是用来预防疾病的）。然而，在寻找原因的时候，我们想了解的是为什么发生的是这件事而不是那件事。

尽管继亚里士多德之后还出现了其他里程碑式的成就（比如 13 世纪阿奎奈的贡献），然而真正的巨大飞跃却发生在文艺复兴末期的科学革命时期。在这一时期，伽利略、牛顿、洛克等人都取得了巨大成就，但是真正为当今因果关系思维和寻找因果关系的方法论奠定基础的是 18 世纪的大卫·休谟。[6]这并不是说休谟做的一切都是对的，也不是说所有人都赞成他的观点甚至信他所信，而是说他以一种批判性的方式重新定义了这个问题。

休谟不单单提出了“是什么使得某事成为了原因”，而是将这一问题一分为二：**何为原因？如何才能找到原因？**更重要的是，休谟没有去寻找能够区别原因与非原因的特征，而是从本质上将二者的关系提炼成了经常性事件。也就是说，我们通过经常性地观察事件发生的规律来了解因果关系，而且也只能通过经历这些有规律的事件来了解原因。

蚊虫叮咬是传染疟疾的必要前提，但春季冰淇淋小贩的突增却不是天气变暖的必要前提。然而，我们无法仅通过观察就找出经常性事件（天气/冰淇淋小贩）与必要性事件（蚊子/疟疾）之间的差异。只有在出现反例时，比如天气已经变暖了，而冰淇淋摊位却并没有增加，我们才能了解到冰淇淋小贩并不是气温变化的必要条件。

我们想当然地认为原因发生在结果之前，而不是在结果之后或与结果同时发生。这一点我们会在第 4 章借用物理学中同时性因果关系的例子来进一步讨论。此外，我们还需关注一些原因并没有在结果之前发生的情况。具体来说，我们所观察到的事件发生的时间也许并不忠于实际上事件发生的时间或事物之间的联系。开枪时，我们先看到的是枪火，然后才会听到巨大的响声。因为我们总是先看到枪火，再听到枪声，所以可能会认为是枪火引起了枪声，但实际上枪火和枪声都是开枪引起的。只有研究了这两个事件发生的共同原因，我们才能理解这种规律性。

很多情况下，我们可能无法在事件真正发生之时对其进行观察，所以即使有些事件其实是有先后顺序的，但它们看起来也好像是同时发生的。这种情况经常出现在病历中：病人诉说一系列症状，然后医生将这些症状记在相应的药物旁边。看起来似乎症状、诊断和处方是同时发生的（因为它们都是在看医生的时候被记录的），即便药物是在症状出现之前服用的（正是因为用药出现了症状才去看医生的）。时间也有可能是错误的，因为数据并不是在事件发生时收集的，而是事后收集的。如果我问你上次头疼是什么时候，除非你专门做了记录，或者是你最近才头疼过并且记忆

犹新，否则你回答的时间可能并不是真正准确的时间。而且事件发生的时间越久，你的记忆就越不可靠。[7] 然而，要想判断一种药物是否真的有副作用，事件发生的先后顺序是最关键的信息之一。

休谟不仅要求原因在时间上早于结果，还要求原因和结果在时间和空间上的距离都要相近（相邻）。如果它们在时间和空间上差得太远，那我们将很难发现它们之间的因果关系，因为很多其他因素可能会掺杂其中并对结果产生影响。假设一个朋友借用了你家的咖啡机，在她归还后的第三个月你发现机器坏了，这时你就很难将责任归咎于你的朋友。但如果她归还机器的时候你就发现机器坏了，那就很容易将责任归咎于她了（事实上，心理学实验也通过让人们根据两个事件发生的不同时间差来推理因果关系，证实了这一现象[8]）。同样，如果一个人距离书架几英尺远，而另一个人离书架的距离比他要近得多，这时一本书从书架上掉下来了，那么站得近的那个人更可能是引起书本掉落的原因。类似地，台球杆击中台球之后，台球立即开始在球桌上滚动，这使得台球与球杆的联系明显多了。

休谟要求原因和结果在时空上具有邻近性，然而有些因果关系却并不符合这一要求。这就限制了该理论的适用范围以及我们进行因果推理的能力。比如说，某种因素的缺乏会导致某种结果，就像缺乏维生素 C 会导致坏血病。这一因果关系就不符合休谟的邻近性要求。如果心理状态（比如信念或意图）也能作为原因的话，那么我们就又得到了一种因果关系，这种因果关系的因果之间没有任何物理上的联系。比如说，学生做作业可能是为了得高分，但是这种得高分的欲望和做作业的行为之间并没有物理上的联系。还有一些时间跨度很长的因果事件，比如因环境因素而导致的健康问题。有时即使这些事件之间是有紧密联系的，我们可能也不会注意到这些联系。[9]

按照休谟的理论，如果我们多次在看到有人按蜂鸣器之后听到声响（经常性联系），就会由此推断按蜂鸣器会导致这种声响。之所以如此推断，

是因为我们看到人的手指接触到了（空间邻近性）按钮，而接触按钮的行为发生在声响之前（时序性），而且在手指接触按钮后几乎立即（时间邻近性）产生了声响。相反，如果这两件事之间有很长的延迟，或者这两件事同时发生，或者蜂鸣器并不是每次都会发出声响，那我们就不能做此推断了。我们不能说按下按钮是发出声响的必要条件，只能说我们多次看到这一事件。关于这方面的知识还有很多，我们将在第 5 章详细讨论。在此引用这个案例主要是为了区分：(1) 产生某种结果的必要条件和伴随条件；(2) 事物之间的潜在关系是什么，以及我们能够通过观察学到些什么。

值得注意的是，并不是所有人都赞成休谟的观点，尤其是康德。众所周知，康德不赞成休谟把因果关系简化为规律，他认为必然性是因果关系的基本特征，而且由于我们无法凭经验推理出事物之间的必然联系，也就无法通过观察归纳出事件发生的原因。相反，他认为我们可以用一种先验知识去阐释我们所观察到的因果关系。[10]

尽管大部分有关因果关系的定义都是基于休谟的理论建立的，但是没有任何一个定义能够包含所有可能出现的情况，每一个定义都有其他定义所没有的例外情况。比如说，某种药物可能只会在个别患者身上出现副作用（所以我们不能假定某个原因必然会产生某种结果）；安全带一般可以防止交通事故中的死亡事件，但在有些情况下却可能会引发死亡事件（所以我们需要想到有些因素在不同环境下可能会产生不同的结果）。

这个问题常被归结为：我们应该将原因视为这个世界的基石或原始力量（这种东西无法简化为任何定律），还是我们强加给事物的一种结构？人们对因果关系的方方面面都存在不同的见解，这一问题也不例外（人们甚至对“某些特定的理论是否能与因果实在论的概念兼容”也各执己见）。有些人认为，原因如此难找，我们根本不可能找到，甚至觉得那些物理学

定律都比原因有用得多。也就是说，他们认为“原因”只是“引发”“推动”“抵制”“阻止”这类词的简约表达，而不是一个基本的概念。[11]

因果关系在日常生活如此重要，但在哲学上却没有一个公认的关于因果关系的理论，也没有什么万无一失的计算方法能帮助我们准确找到因果关系，这让人有点惊讶。但更棘手的是，由于人们对“原因”的定义不同，所以同一情况下，人们可能会将不同的因素视为事件发生的原因，但事件的真相可能并没有人知道。

比如说，鲍勃遭遇了抢劫，而且劫匪想要杀人灭口。但在抢劫的过程中，鲍勃心脏病突发，随后死亡。我们可以将鲍勃的死因归咎于生理机制（心脏病发作），并进一步追溯到心脏病的根源——遗传基因，这种基因大大增加了心脏病突发致死的概率。或者将鲍勃的死因归咎于抢劫事件，因为如果没有遭遇抢劫，鲍勃的心脏病就不会发作。这两种死因都解释得通，我们无法立即搞清楚哪个解释更合理，或者它们只是对一个事件的两种分析。此外，不要试图为事件找出某个唯一的原因。也许是心脏病发作和抢劫事件共同导致了鲍勃的死亡，这两个事件的影响是不可分割的。在第 8 章和第 9 章中，我们将再次分析这两件事对鲍勃的死亡所应承担的责任，并研究一些事件发生的原因（比如为什么会爆发某场战争）以及某些政策是否有效（比如禁止在酒吧吸烟的政策是否改善了纽约市的人口健康状况）。

尽管原因不易寻找又难以界定，但也不是毫无希望。答案并不像人们想象的那么清楚明了（我们没有神奇宝盒，不能从这头输入数据然后等它自动输出原因，并且输出绝对正确、万无一失），我们的大部分工作只是找出何时该用何种方法。关于原因的定义有很多不同的观点，这些观点给我们提供了很多种方法，这些方法或多或少都有点用，只是工作原理和适用的情形有所不同。如果能了解其中两种或以上的方法，并且了解它们之间是如何互补的，那么我们就能以多种方法来考察同一种情形了。有些

方法适用的情形可能比较多(或者适用于对我们而言很重要的一些情形),但是请记住，没有哪种方法是十全十美的。尽管寻找原因很难，但一定要坚持不懈地去寻找正确的原因。如果能够坦然接受我们可能会犯错的事实，并且明确在何时能够找到什么，那么我们就可以不断地尝试，看看这些方法都能适用于哪些情形，至少能准确地描述出我们所使用的方法以及所得到的结果。本书重点阐述了各种方法的优势和局限性，而不是向读者推荐某些方法，因为这些方法都不是绝对的。数据不全时可能这种方法更有效，事件发生的时间很重要时可能那种方法更有效，总之，具体使用哪种方法要视情况而定。

因果思维对科学、法律、医学和其他领域(很难想出有哪一个学科不关心或者不需要找到事件发生的原因)都至关重要，但其缺陷之一在于，用来描述原因的语言和用来寻找原因的方法可能过于专业化，并且让人感觉它只局限于特定领域。你可能觉得神经学和经济学之间没什么共同点，也不认为计算机科学能够解决心理学问题，但这些不过是新兴的、跨学科研究因果关系的一部分领域。然而，所有的领域在哲学上的起源都是一样的。

1.2 怎样才能找到原因

哲学家们长期以来一直在关注“原因究竟是什么”这个问题，但是界定因果关系的主要哲学方法以及我们今天用来从数据中寻找因果关系的计算方法，却直到 20 世纪七八十年代才出现。我们不知道将来是否会出现一个公认的因果关系理论，但我们有必要了解这个广泛使用的概念的含义，只有这样才能更清晰地对它进行思考和讨论。我们在因果关系研究领域所取得的任何进步都会对计算机科学以及其他领域产生广泛的影响。假如原因不仅是一种事物，那么我们可能就要用多种方法去寻找它、描述

它，并且用不同的实验来验证人们关于原因的直觉。

自休谟以来，因果关系研究领域所面临的主要问题是：我们该如何区分包含因果关系的事件和不含因果关系的事件。20 世纪六七十年代出现了三种主要的研究方法，都建立在休谟的理论基础之上。单一的原因不太可能引起某种结果，所以 John L. Mackie 提出了一个理论，他认为某种结果的产生是由一系列条件共同导致的。[12]这一理论很好地为我们排除了不包含因果关系的事件，并且解释了原因的复杂性。类似地，许多因果关系都包含偶然性因素，在这类情况下，原因可能只是提高了某种结果出现的可能性，但并不保证它一定会出现。针对这一特征，Patrick Suppes 及其他研究者们提出了概率法。[13] 休谟的理论还促成了反事实推理法：通过假设导致某件事的原因不存在，事情的发展会有何不同，从而来界定这一事件发生的原因。[14]比如说，某个人是赢得一场比赛的主要原因，因为如果没有他，这场比赛就不会赢。

哲学上的这些方法似乎已经脱离了寻找因果关系的计算方法，但这些不同的因果思维却能为我们提供许多方法去寻找因果关系的证据。对于计算机科学家来说，人工智能的梦想之一就是实现自动推理。要做到这一点，关键之一在于找到事件发生的各种原因，并利用它们来形成各种解释。这项工作在现实生活中得到了广泛的应用，从机器人的生产（机器人需要使用现实世界的各种模型来计划自己的行为，并预测这些行为的结果）到广告宣传（亚马逊如果知道你点击“现在购买”按钮的原因，就能向你推荐更适合你的商品）再到医疗服务（重症监护病房里的患者的身体状况突然发生变化时，会向医生发出警报）。然而，要想制定出算法（解决问题的一系列步骤），我们需要对问题进行精准的描述。要想设计出能够找到原因的计算机程序，我们需要对原因进行定义。

20 世纪 80 年代，以 Judea Pearl 为首的计算机科学家们向人们证实了，以概率来定义因果关系的哲学理论可以用图表来表示。这些图表可以向人

们直观地呈现出事件之间的因果关系，并为人们提供了针对不同变量之间的数学关系进行编码的方法。更重要的是，他们还引入了一些根据先验知识来构建图表以及从数据中寻找它们的方法。[15]这就为我们带来了很多新的问题。如果因果事件之间存在可变延迟，那我们还能找到因果关系吗？如果因果关系本身会随着时间而发生改变，那我们能从中学到什么？计算机科学家们设计了一些能够自动寻找事件的解释的方法，以及测试这些解释是否符合实际的方法。尽管我们在过去的几十年里取得了很多成就，但是依然面临着许多挑战，尤其是我们对数据的依赖程度已经越来越高。我们现在所面临的不是那些为了研究而精心挑选出来的数据集，而是海量的、不明确的、根据观察得到的数据。想象我们正面临这样一个简单的问题：根据 Facebook 数据了解人们的人际关系。第一个困难是，并不是所有人都使用 Facebook。所以，我们只能通过 Facebook 研究一部分人的人际关系。这部分人也许并不能代表所有人，也不能代表你感兴趣的某一类人。此外，人们使用 Facebook 的方式也不尽相同。有些人从来不会显示他们的人际关系，有些人可能会显示虚假的人际关系，还有些人可能不会及时更新他们的个人信息。

在因果推理过程中，尚未解决的关键问题包括：从不明确的或缺少变量和未经观察（如果我们没有观察吸烟这个变量，是否会错误地把其他因素当作引起肺癌的原因）的数据中寻找事件的原因，寻找事件之间的复杂关系（如果这个结果是一系列事件共同导致的呢），以及寻找偶发事件的原因和结果（是什么导致了 2010 年股市的闪电崩盘）。

有趣的是，电子健康记录等海量数据正将流行病学与健康计算工作相结合，以了解影响人口健康的因素。我们的研究是先了解影响健康的因素，然后利用这些知识来指导公共健康干预措施，而大量人口的长期健康数据（他们的诊断、症状、用药情况、所接触的环境等）对研究有莫大的帮助。我们面临着双重挑战——研究设计（流行病学的一贯研究重点）

并从大型数据集（计算机科学的主要焦点）中进行高效且准确的推理。由于流行病学的研究目标比较特殊，所以它在设计方法以寻找原因方面有着很长的历史，从 James Lind 随机检查水手来寻找坏血病的病因 [16]，到 John Snow 发现被污染的水泵是导致伦敦霍乱疫情的一个原因 [17]，到 Koch 提出的假设在细菌和肺结核之间建立了因果关系 [18]，再到 Austin Bradford Hill 将吸烟和肺癌联系在了一起，并为人们评估因果关系提供了一些指导原则。[19]

医学研究也比以前更加依赖数据了。各大医院和私人诊所都在将病人的病历从纸质图表转换为电子格式，但这种转换工作必须满足有意义的使用标准（比如能够利用数据来帮助医生诊断病情），它所带来的好处要能够抵消转换工作所消耗的成本。然而要想满足这些标准，很多工作都要进行海量的数据分析，这就需要使用计算方法。

神经科学家可以通过脑电图描记器和功能磁共振成像仪来收集有关大脑活动的海量数据，并利用计算机科学和经济学的研究方法来分析这些数据。脑电图中的数据本质上就是大脑活动的量化数字记录，这种记录在结构上和股市数据差不多（股市数据可以告诉我们随着时间的变化，股票的交易价格和交易量是多少）。Clive Granger 提出了经济时间序列中的因果关系理论（他因此获得了诺贝尔奖），这一理论不仅适用于经济学，还被应用于其他生物学数据，如基因表达阵列（用来测量随着时间的变化，基因的活跃程度如何）。[20]

经济学中的一个关键挑战是，判断执行某个政策是否能实现预期的目标。这与公共健康领域所关注的问题十分类似，比如判断是否可以通过减少苏打水的瓶身容量来减轻肥胖症问题。这个问题也是我们所面临的最难解决的问题之一。在很多情况下，所颁布的政策本身就会改变社会的体制。我们会在第 9 章看到这样一个例子：田纳西州最初做了一个缩小班级规模的实验，于是加州用一种十分仓促的方式也缩小了班级规模，但这两

个事件的结果截然不同。如果所有条件都不变的话，那么一项干预政策可能会带来积极的影响，但也可能会改变人们的行为。如果要求人们系安全带的法规会导致人们开车时更加鲁莽，那么我们就很难了解这个法规的影响究竟是好是坏，以及在交通事故死亡率不降反升的情况下，到底是要废除这一法规还是进一步完善它。

对于心理学家来说，理解因果推理（包括它的发展过程，人与动物之间的差异，以及它何时会出错）是理解人类行为的关键之一。经济学家也想知道人们为什么会做出各种行为，尤其是在做决策的过程中。最近，心理学家和哲学家共同利用实验方法来研究人们对因果关系的直觉（这属于实验哲学的研究范畴[21]）。一个很关键的问题在于，要理顺因果关系和道德评判之间的关系。如果有人在资助申请中杜撰数据并因此获得了资助，而其他诚实可敬的科学家们却因为资助资金有限而没有获得资助，那么我们能说是那个欺骗者导致他们没有获得资助吗？现在有两个问题：应该怪罪那个欺骗者吗？如果所有人都存在欺骗行为，那么我们对这件事的看法是否会发生改变呢？要了解人们是如何做出因果关系判断的，这不仅能帮助我们更好地理解人们的思维方式，还能帮助我们处理一些实际问题，比如解决分歧、提升教育和培训水平[22]以及保证陪审团的公正性。本书会告诉大家，虽然我们无法消除所有导致偏见和错误的因素，但可以更准确地发现这些因素并了解它们可能会带来的影响。

1.3　为什么需要原因

原因难以界定又不易寻找，那么它们对我们究竟有什么好处呢？我们又为什么需要它们呢？有三件很重要的事只有在清楚原因的情况下才能做到，或者做到最好，这三件事是：预测、解释和干预。

首先，假设我们想要预测谁会赢得美国总统大选。专家们找到了各

种规律，比如共和党人必须赢得俄亥俄州的选票才能赢得大选；自富兰克林·罗斯福之后，没有任何一位总统能够在失业率超过 7.2%[23] 时获得连任；美国从来没有女性总统（至少在我写作本书时是这样的）。[24] 然而这些只是规律而已。我们可以在历任总统身上找到很多这样的规律，但是我们无法从中得知他们**为什么**会赢得大选。人们是根据失业率投票的吗？还是说失业率只是间接反映了国家形势和经济状况，暗示人们在失业率高的时候要寻求变革？更糟糕的是，如果我们发现的这些规律只是巧合，那么它们最终都会被打破。而且，这些数据是从很小的数据集中得出的；美国历史上只有 44 位总统，其中连任的总统还不到一半。

这就是黑盒问题：我们把数据输入黑盒子，然后从中得出一些预测，但是黑盒子不会对这些预测做出任何解释，也不会告诉我们这些预测为什么值得信赖。如果我们不知道这些预测为何会成真（为什么赢得某个州的选票就能赢得大选），也就无法预料它们的失败。如果我们知道俄亥俄州能够“决定”一场大选的原因是这个州的人口特征十分具有代表性，而且这个州从来不专属于某一个政党，那么我们就能由此预测。如果由于移民人口导致俄亥俄州的人口组成发生了巨大变化，那么之前的预测——它对大选有决定性作用——也就不复存在了。如果这个州只是反映全国总体趋势的一个间接指标，那还可以通过全国民意调查来获得更直接、更准确的预测。一般来说，与相关性相比，原因能够为我们提供更为可靠的方法来预测事件的结果。

再举一个例子，比如说某种基因的变异导致了运动耐量的提高和免疫反应的增强。然后我们可能会发现，运动耐量的提高对人体免疫反应来说是个好指标。然而，运动耐量的高低变化只是一个非常粗略的估计，因为除了基因突变以外，还有很多其他因素（比如充血性心力衰竭）也会导致运动耐量的变化。因此，只根据运动耐量进行诊断可能会导致很多误诊，错误地夸大或低估病人的病情。更重要的是，一旦了解到基因变异会引起

运动耐量和免疫反应的双重提高/增强，我们就能获得两种测算风险的方法，并且能够避免收集过多的测量数据。既然运动耐量只是反映了基因的变化，那么我们就没有必要对这两者都进行测试。但值得注意的是，如果基因测试极易出错的话，那么运动耐量的测试数据也许能为我们提供确凿的证据。还有一点，将患者送到运动生理学实验室去测试他的运动耐量的成本，可能要比单独测试某一种基因变体高得多。然而，我们无法将测试方法的直接性和它所花费的成本进行比较（如果运动耐量测试的成本比基因测试低得多，那我们可能更倾向于先测试运动耐量，尽管这种测试方法是间接的），除非我们知道这些因素之间潜在的因果关系。因此，即便我们只想预测谁会赢得大选，或者某个病人患某种疾病的风险有多大，只要了解了那些因素为什么具有预测作用，就能够提高决策的准确性并降低决策的成本。

现在，我们想知道为什么有些事件是相互关联的。视力模糊和体重下降之间有什么联系？如果只知道这两个症状经常相伴出现，是无法得出更多信息的。只有找到导致这二者的共同原因——糖尿病——我们才能理解它们之间的关联。很明显，在这类事件中，我们要找到事情发生的原因，而这也是我们一直在做却极少深入研究的事。

也许你曾看过有关“食用红肉的人群死亡率更高”的研究，但如果你不知道其中的原因，那这些信息就是不可用的。也许吃红肉的人喜欢饮酒或不爱运动，这些都是影响死亡率的因素。而且，即便死亡率的升高真的是红肉引起的，与其他因素无关，那也要根据具体情况来决定用何种方法来降低这种风险。如果死亡率的升高是不卫生的烧烤方式造成的，那我们可以换一种烹调方法；如果是吃红肉本身引起的，那我们就只能让自己成为素食主义者了。我们想知道的不仅是红肉是否与死亡率有关，而是红肉是否真的会提升死亡率。我之所以格外强调这种说法，是因为报纸的科学版块几乎每周都会写一些与饮食和健康相关的内容，比如鸡蛋能引发/

预防各种病痛，咖啡会提高/降低死亡的风险。这类研究有时可能不仅会提供某件事与某类人群之间的相关性，还会提供一些其他证据，但是，所有这类研究都值得怀疑，我们要对每一个细节进行批判性的考证，尤其是要用这些结论来指导各种政策和行为的时候（第 9 章将进行讨论）。

有些时候，我们要去解释一些事件发生的原因。你上班为什么迟到了？某人为什么生病了？为什么一个国家入侵了另一个国家？在这些情况下，我们想知道是谁或者是什么因素引发了某个事件。迟到与交通有关；随着年龄的增长，人们会患各种疾病；很多战争都是由于人们在意识形态上存在分歧。但这些并不能告诉我们上述事件发生的原因。你迟到可能是因为汽车抛锚了，Jane 生病可能是因为食物中毒，某场战争可能是领土或资源争端引起的。

找到事件发生的根源很重要，它不但会影响政策的制定（如果 Jane 生病的原因是餐厅的卫生条件太差了，那她可以不再去那家餐厅吃饭，但无须避讳那天所吃的食材）与责任的归属（谁该为 Jane 的病情负责），还会影响人们对某件事的反应。很多疾病的症状可能与服用治疗该疾病的药物后产生的症状相同。比如说，慢性肾病会导致肾衰竭，但在极少数案例中，治疗慢性肾病的药物可能会对肾造成同样的损伤。如果门诊医生看到患有肾病的人同时也在服用会导致肾损伤的药物，那他就需要明确这个病人的肾病是否是由服用的药物导致的，这样才能为病人制定正确的治疗方案。虽然知道了服用某些治疗肾病的药物也可能导致肾损伤，但医生无法仅根据这一点就确认某个病人是否属于这种情况。只有在确认病人是否属于这种情况后，才能决定是否要让病人停止服用这种药物。

因果关系最重要的用途是可以用来干预某些事情的发生。我们不仅想知道某些事件为什么会发生，更想利用这些信息来预防或促成某些结果的产生。你可能想知道如何通过改变饮食习惯来改善身体的健康状况。需要服用维生素吗？要坚持吃素吗？还是要戒掉含碳水化合物的食物？如

果这些干预措施并不能带来你想要的结果，那就没必要做这些费时又费力的改变。同时，我们还需要考虑这些干预措施的成效如何。也许你听说某个节食方案的减肥成功率是 100%。在基于这句话做出任何决策之前，你应该先了解一下这个节食方案究竟帮助人们减掉了多少体重，不同的人减掉的体重差是多少，这个节食方案的效果和其他节食方案相比有何差别（仅通过自己有意识地控制饮食也是可以减肥的）。我们既要评估已经采取的干预措施是否有效（纽约市在发布食物的卡路里值后，是否改善了市民的健康状况），也要预测将来可能会采取的干预措施的效果（如果减少快餐中的钠含量，会出现什么情况）。

政府部门必须知道他们的政策会对民众产生什么样的影响，并且必须制定出能满足民众需求的政策。比如，研究人员发现含钠量高的食物与肥胖症有关联。于是，立法人员决定颁布一项法案，旨在减少餐厅食物和包装食品的含钠量。如果含钠量和肥胖症之间的唯一联系是，高热量的快餐食品导致了肥胖症，而这些食品又正好含钠量高，那么这项法案将不会产生任何作用。人们依然会购买快餐，而快餐才是一开始就应该关注的问题。我们必须保证我们的干预措施针对的是真正影响结果的原因。如果我们只干预了一些与结果相关的因素（比如通过禁止使用火柴来减少死于因吸烟导致的肺癌的人数），那这样的干预措施是不会有效果的。

如果干预措施还有副作用的话，那么情况就更加复杂了，这一点我们后面再讨论。因此，我们不仅要知道造成某个结果的原因，还要知道这个结果会带来什么影响。比如，增加运动量会导致体重下降，但是“补偿效应”又会导致人们去摄入更多的热量，甚至比他们消耗掉的热量还要多（于是他们的体重不降反升）。所以我们要做的不是去寻找个体变量之间的单一联系，而是要对事物间各种相互关联的关系有一个更为宏观的认识。

1.4　接下来……

人们为什么会在不相关的事件之间看到关联性？陪审团如何评估犯罪的原因？我们如何通过实验来得知某个病人应该服用哪种药物？随着我们对数据和算法的依赖程度越来越高，了解因果关系已经成为一项必须掌握的技能。我们不仅需要利用这一技能从数据中提取有用的信息，还要用它来指导日常生活中的各种决策。即使你的工作并不包括做研究或分析数据，因果推理的各种潜在用途也会对你产生影响，比如你要与别人分享什么样的个人信息，以及与哪些人分享。

为了更加准确地寻找和使用原因，我们需要知道因果推理过程中的心理活动（我们是如何感知并推理事件发生的原因的），还要知道如何评估我们手中的数据（不管是通过观察还是实验获得的），以及如何利用这些知识进行决策。尤其是要考察所收集的数据（以及我们操控这些数据的方式）是如何影响我们从中得出的结论的。在本书中，我们将探索如何利用各种论据来支持或反对某种因果关系（既是正方也是反方）、如何利用因果关系中的信号来超越那些间接的证据，以及如何准确地找到并理解这些信号。

注释

1. 想要了解 Meadow 所用的数据，参见 Fleming 等（2000）。想要了解 Meadow 对使用这一数据的评论，参见 Meadow（2002）。
2. Meadow 因为在证词中使用了这一数据，后来被判渎职罪，并被吊销了医生执业资格，导致他不能再行医（后来他通过上诉得以恢复执业资格）。
3. 参见 Thompson 和 Schumann（1987）。还有一个著名的案例是 Lucia de Berk 案。Lucia de Berk 是荷兰的一名护士。像 Clark 一样，她一开始也被误判为有罪，后来又被宣布为无罪。De Berk 护理过很多意外死亡的病人，一名专家证人计算了这一情况发生的概率，得出的结果是这一情况完全是巧合的概率只有三亿四千二百万分之一。了解更多关于 Lucia de Berk 案的信息，请参见 Buchanan（2007）。正如 Clark 案一样，这个数据被等价为 De Berk

无罪的概率。检方认为这一概率发生的可能性如此之小，所以它一定不可能发生。

4. 值得注意的是，SIDS 只是导致婴儿猝死的原因之一。事实上，在 Clark 一案中，有重要证据表明其中有一个婴儿患有炎症，并且这一炎症可能会对婴儿造成生命危险。然而，参与该案的病理学家（该病理学家后来被判严重渎职罪，并被禁止行医三年）在庭审中并未公布这一证据。
5. 亚里士多德关于因果关系的论述，参见亚里士多德（1924，1936）。想要了解古希腊学者关于因果关系的论述，参见 Broadie（2009）。
6. 休谟（1739，1748）。
7. 人们关于时间的记忆具有两个特征：不确定性和特异性，关于这两者之间的关系，参见 Hripcsak 等（2009）。
8. 具体案例参见 Lagnado 和 Speekenbrink（2010）。
9. 注意：休谟一定不会赞成这样的评价。他认为如果原因和结果之间存在时间或空间上的间隔，那人们就会发现一系列将原因和结果连接在一起的中间原因。
10. 参见 Kant（1902，1998）。
11. 参见 Cartwright（1999，2004）和 Skyrms（1984）。
12. Mackie（1974）。
13. Suppes（1970）。
14. Lewis（1973）。
15. 关于这些内容的专业性介绍，参见 Pearl（2000）和 Spirtes 等（2000）。
16. Lind（1757）。
17. Snow（1855）。
18. Koch（1932）。
19. Hill（1965）。
20. Granger（1980）。
21. 想要了解更多关于实验哲学的信息，参见 Alexander（2012）、Knobe 和 Nichols（2008）。
22. 当判断因果关系的过程中存在多种文化差异时，情况更是如此。比如说，有些人可能会将技能看成是一种天生的能力，人们要么有这样的技能，要么没有，而其他人可能会认为根据环境和人们的努力程度，一个人的技能是可以改变的。
23. Appelbaum（2011）。
24. 有一幅很棒的漫画向我们阐释了各种主观的规律，名叫“Electoral Precedent”。

第 2 章　心理

人们是如何寻找原因的？

1692 年，马萨诸塞州塞勒姆镇上有两个小姑娘突然行为失常。Abigail Williams（11 岁）和 Elizabeth Parris（9 岁）突然出现了痉挛和抽搐的症状。由于她们没有任何明显的生理疾病，于是医生认为她们的古怪行为可能是巫术造成的。不久，又有几个小女孩出现了同样的症状。接着，有十几个人因此被指控为女巫。

人们一直认为塞勒姆镇的女巫审判案是一场大规模的癔症和骗局，但在近三百年之后却提出了一个新的假设：麦角中毒。[1] 食用麦角菌（生长在黑麦和其他谷物上的一种菌类）会导致麦角中毒—— 一种会出现癫痫、瘙痒症状，甚至会影响精神的疾病。这一假设利用当时的天气记录来说明当时的环境很适宜麦角菌的生长，而且女巫案发生的时间也正是收获与食用黑麦的季节。不过，这似乎暗示了还有很多人应该也吃了黑麦，却没有出现麦角中毒的症状（这就降低了这个假设的可信度），但因为孩子们可能更容易受到麦角中毒的影响，所以可能只有他们出现了中毒的症状。此外，还有一位历史学家发现发生女巫案的区域、黑麦的价格及收获季节这几个因素之间存在相关性。[2]

麦角中毒似乎是一个非常合理的解释，但有些证据却是相互矛盾的。麦角菌能够引起两种中毒症状——坏疽症和抽搐症，但塞勒姆镇没有关于

暴发坏疽症的记录。而且，这种抽搐症很可能会影响所有家庭成员，人们也曾认为这是一种传染病。[3] 越小的孩子越容易感染这种疾病，但在塞勒姆女巫案中，出现这种症状的绝大部分都是十岁以上的孩子。然而最离奇的是，这些症状似乎会因为“女巫”的出现而受到影响，因为这些女孩一离开法庭，身上的症状就会减轻。如果这些症状是由麦角中毒引起的，那么似乎就不应该因为在场人员的变化而发生如此戏剧性的改变。

尽管麦角中毒的解释遭到了反驳[4]，但一直到 1982 年，《纽约时报》都还在发表有关麦角中毒的文章。[5] 不论何时何地，人们都愿意去相信那些符合他们当下认知的因果解释，即使这些解释与所得数据并不吻合。在 17 世纪，人们认为巫术是很合理的解释，并会重点强调支持这一解释的事实，尽管这些事实是带有高度偏见的、不科学的试验，如“幽灵证据”（原告看到被告伤害他们的幻象）等。在 20 世纪，像“中毒”这样的科学解释更加容易理解，但还是无法解释为何这些症状只出现在一群十几岁的孩子身上。

---

17 世纪初，人们之所以会认为巫术是一种合理的解释，是因为他们对原因的认知是由他们对现实的感知、基于经验的推理以及已有的知识组成的。物理学告诉我们：如果你击打一个球，它就会开始滚动。但如果你之前了解到的知识是地球是一个平面，或者巫术能把物体从房间的这头移动到那头，那么你可能就会对台球的运动原理做出不同的预测和解释。

知道在哪里更容易找到原因、哪里更容易出错既有助于我们设计出更好的数据分析软件，也能对日常生活有所帮助。本章将探讨我们对因果关系的认知是如何随着时间的变化而发生变化的，以及我们是如何通过对世界的观察和与世界的互动来把握事件发生的原因的。当我们想要评判一个人的行为时，比如责怪某人害我们上班迟到了，或者表扬某人谨慎的开

车态度，我们的推理就不仅基于因果关系了。明确哪些其他因素（比如期待）会影响我们对责任的判断，有助于我们更好地理解这一行为。然而，人们对于事件发生的原因（比如赢得一场比赛）可能会有不同的看法。我们从一个人群中学到的判断因果关系的知识也许并不适用于另一个人群。所以，我们将会研究影响因果关系判断的社会因素和文化因素。最后，还会讨论为什么我们如此容易受到因果谬论的影响，以及为什么我们明知会被错误的因果关系观念（比如迷信）影响，但它们还依然存在。

2.1　原因的寻找与使用

你是怎么发现按一下开关灯就会亮的？你是怎么知道是先开枪然后发出的声音，而不是先有声音后开枪的？因果关系的学习主要包括两点：感知（对因果关系的直接体验）和推理（从不含因果关系的信息中进行推断）。

当我们在感知因果关系时，并不是要通过模式识别的方式将我们所观察到的内容与先前的知识相联系，而是要去亲身体验这种关系。当看到一块砖头飞进窗户、一个台球被另一个台球撞击后开始滚动，或者一根火柴点燃了蜡烛，你就会根据这些感官输入而感觉到事件间的因果关系。相反，像食物中毒、战争和身体健康等现象就无法通过观察而直接感知到它们发生的原因，必须通过其他方法来进行推理。

“我们能够感知到因果关系”这一观念在哲学领域是有争议的，而且与休谟的理论正好相反，休谟认为我们只能通过观察来了解事物间的因果关系。本章将会展示一些强有力的实验证据，以此来证明我们是能够感知到事件间的因果关系的。感知理论认为，人的大脑中存在某种程序，可以接收外界输入的信息并将这些信息分成有因果关系的和没有因果关系的，而不是通过其他线索来寻找事件发生的原因。尽管心理学研究已经证明大脑具有感知因果关系的能力，但仍然存在一个问题：推理和感知是否是

大脑中可分离的两种活动？为了验证这一问题，人们利用一些案例来进行实验，这些案例中的感知和判断是相互矛盾的，因为如果感知和判断是同一种活动，那么两种情况下的答案应该是一样的。研究结果证明，人们在相互矛盾的感知和判断案例中给出的答案确实是不同的。但因为这些答案是人们对自己直觉的描述，所以我们是不可能完全将感知从推理中分离出来的。[6]

要想设计出能够将这两种活动分开的实验（要确保判断活动没有感知的参与，感知活动也没有判断的参与）并不容易，但是针对裂脑患者的研究却为我们提供了一些线索。这些患者的大脑左右半球之间的联系已经部分或完全断裂，所以两个脑半球之间的任何信息传输都有延迟。这对感知的研究十分有帮助：如果感知和推理是由不同的脑半球来处理的，那么这些患者的大脑就有可能独立地呈现这两种活动。研究人员通过每次只在某个视线范围内展示刺激物，从而控制大脑的哪个半球能够接收到信息。在感知和推理因果关系时，正常人的大脑并没有表现出任何差异，但是裂脑患者却有显著的差别，这是因为执行这项任务的是不同的大脑半球。这似乎表明了推理和感知是可以分离的两种活动，并且这两种活动所用到的大脑区域可能是不同的。[7]

2.1.1 感知

这些研究表明，感知活动可以独立于推理活动而发生，但我们究竟什么时候才能感知到事物间的因果关系呢？Albert Michotte 对感知因果关系的基础性研究向我们证实了这一点：当人们看到图像中的一个模型向另一个模型运动并且击中了它，然后第二个模型开始运动，他们就会感觉第二个模型是由第一个模型“发动”的。[8]尽管这只是一些图像而不是实物，但是这个结论却非常真实，还有其他研究者也做了相同的实验并得出了同样的结论。Michotte 的研究为因果关系心理学奠定了基础，同时，他所设

计的在事件间存在不同时间延迟和间隔的实验，也为我们提供了很多关于时间如何影响感知的真知灼见，这些内容我们将在第 4 章进行讨论。

我们对因果关系的理解是如何演变的？这其中又有多少是后天习得的？解答这些问题的关键之一来自我们对婴儿的研究。如果我们能够直接感知到因果关系，那么婴儿应该也有这种能力。当然，测试婴儿对因果关系的感知能力不是一件容易的事，因为我们无法询问这些婴儿的感受。

有证据表明婴儿看新事物的时间要长一些，因此研究人员让这些婴儿先熟悉某个事件序列，然而再将这一事件序列倒过来给他们看，并比较婴儿看这两种事件序列所用的时间。婴儿观看的是一些发动序列的视频（第 4 章将进行详细论述），这和一个台球击中另一个原本静止的台球的事件序列相似。第一个球将动力传给第二个球，然后第二个球按照第一个球运动的方向开始运动。首先按照正常顺序播放这些视频，然后倒序播放（击球的过程反过来了，就好像第二个球击中了第一个球一样）。研究人员还给这些婴儿以正序和倒序播放了一些与上述序列类似但不含发动过程的序列（比如两个模型都朝同一个方向运动，但没有任何接触）。这个实验的主要发现是，婴儿观看倒序序列的时间更长。这两种序列都按照正序和倒序进行了播放，所以如果婴儿在含有因果关系的序列中没有感知到因果关系发生的变化（也就是因果关系的对调），而不含因果关系序列中又没有这种变化，那么婴儿观看这两个序列（逆向播放的含因果关系的序列和逆向播放的不含因果关系的序列）的时间就应该没有差别。[9]

尽管我们似乎在婴儿时期就能够感知因果关系了，但还有一些研究表明，6 个月大的婴儿和 10 个月大的婴儿对更加复杂的因果关系（比如台球的随意碰撞）的感知能力是有差别的。[10]有研究表明，感知能力随着年龄的增长而增强。6 到 10 个月大的婴儿能够感知两个物体之间的简单的因果关系，但因果链实验（诸如绿球撞击红球，红球再撞击蓝球这样的因果关系序列）表明，15 个月大的婴儿和成年人能够感知到因果链中的

因果关系，而 10 个月大的婴儿却不能。[11] 要对比年龄较大的孩子和成人的感知能力是一项十分具有挑战性的工作，因为他们之间的差异可能是由表达能力造成的。有些研究简化了任务，用一套具有某种限制的图像来测试 3 到 9 岁儿童的反应。结果发现，连最小的参与者都具有高级的因果关系推理能力，不过这种能力在不同年龄的儿童身上依然存在差异。[12]

不同年龄的人在感知因果关系的能力上所表现出的最大差异，似乎出现在感知与推理相互矛盾的时候，因为孩子们更依赖于对事件的感性认知，而成年人则更依赖于对事件的进一步了解。有这样一个实验，分别把两个机械装置（一个快一个慢）藏在同一个盒子里，每个机械装置都能敲响铃铛。往盒子里放一个球时，如果使用的是快速机械装置，会立即敲响铃铛，而如果使用的是慢速机械装置，铃铛过一会儿才会响。在参与者熟悉了这两个机械装置并且知道盒子里是哪一个装置的情况下，5 岁的儿童依然是去感知而非推理因果关系，9 到 10 岁的儿童及成人能够正确地推理出因果关系，而 7 岁的儿童则介于两者之间（判断的准确率在 50%左右）。当盒子里是慢速机械装置时，先放进一个球，过一会儿会再放进去一个球。在第二个球进入盒子之后铃铛立即响起，但由于机械装置的延迟，铃铛的响声与第二个球没有任何关系。尽管第二个球不可能导致铃铛发出声响，但是年龄小的孩子依然选择第二个球作为铃铛响的原因。[13]

自 Michotte 以来，许多研究感知的实验都会直接询问参与者对于某个场景的观点，比如让他们描述所观察到的内容。然而，这并不能捕捉到感知过程中的本能反应。为了了解这些反应，研究人员最近在成年参与者身上使用了眼动追踪技术。他们不再记录参与者的观察时长，而是关注他们都观察了哪些地方。这一研究表明，在发动式的事件序列中，参与者会预测由因果关系所导致的运动，并且相应地转移自己的注意力。[14] 这意味着不论参与者是否承认序列中包含因果关系，他们对事件发展的期望都表明，他们预测到一个物体的移动是通过与另一个物体的接触而产生的。在

这之后的另一项研究不仅记录了眼球的运动，还记录了参与者对因果关系的判断（这与 Michotte 的研究一样）。这项研究发现，在简单的事件序列中，眼球运动和因果关系是相关的，但在加入时间延迟之后，参与者的眼球运动和因果关系判断就不再具有相关性了。[15]

在情景简单的实验研究中，产生感知偏差的主要是儿童，但成年人对自身因果关系感知能力的信任也可能会导致他们做出错误的判断。如果你听到一声巨响，然后看到房间里的灯灭了，你就可能认为这两个事件是有联系的，但其实是有人在发生巨响的时候正好关了灯。事件发生的时间以及空间上的邻近性等因素会导致人们错误地感知因果关系，从而做出错误的因果关系判断。我们常听说有人在打了流感疫苗的当天就出现了类似流感的症状，于是就有人认为是流感疫苗引起了这些症状。在前面的实验中，盒子里的慢速机械装置是无法在第二个球进入盒子时就立即敲响铃铛的，同理，流感疫苗里含的是一种不活跃的病毒，这种病毒是无法引起流感的。注射流感疫苗的人有很多，其中有些人可能偶然感染了类似的疾病，甚至有可能是在候诊室里接触到了流感病毒。可以通过关注事件的背景信息，了解所有可能出现的情况，从而避免错误的判断。

2.1.2　推理与论证

当试图搞清楚你的车为什么会发出奇怪的噪声，或者推断傍晚喝的咖啡导致你晚上睡不着觉时，你无法直接感知到汽车热度和刹车发出的噪声之间的关系，也无法直接感知到咖啡中的兴奋因子是如何影响神经系统的。相反，你需要用到另外两种类型的信息：关于刹车系统工作原理的机械知识，以及食用含兴奋因子的食物后该如何入睡。所以，即便我们完全不了解某个原因的作用原理，也可以通过观察原因和结果共同出现的频率来获取一些认知。即使我们观察到的因果关系案例只有一个，也可以根据我们对系统本身的理解来进行推理。因此，我们可以通过对以下两个问题

的理解来推断汽车发出噪声的原因：汽车的各个部件是如何相互作用的，以及系统中的哪些故障会导致这种噪声。有两种推理方法是互补的：一种是利用协同变化法（事件共同发生的频率），另一种是运用机械知识（某个原因是如何引起某种结果的）。尽管研究人员通常将这两种推理方法分开研究，但我们可以同时使用它们来推理事件发生的原因。[16]这种使用间接信息来寻找原因的过程叫作因果推理。进行因果推理的方法有很多，但重点是我们并不能直接感受到因果关系，而是要通过数据和背景知识来推断因果关系。

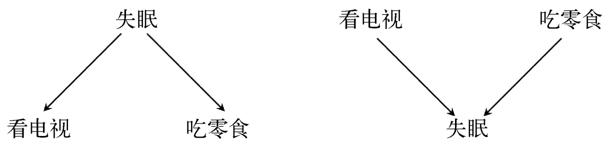
心理学有一个经典的因果推理任务：给参与者展示一系列事件，然后让他们说出是什么原因导致了某种结果（比如某种声音或者屏幕上的某种视觉效果）。其中最简单的任务是，让参与者判断是否（或者在多大程度上）是某一个事件导致了另一个事件，比如让参与者通过一系列观察来判断是否是某一个开关点亮了某一盏灯。研究人员试图通过改变不同的变量（比如原因和结果之间的时间延迟、参与者是否与系统进行互动、因果关系的强度，等等）来破译影响人们推理因果关系的因素。我们已经知道了时空上的距离会让人们觉得某个事物不大可能是引发某起事件的原因，但实际情况并非这么简单。在考察时间如何影响我们对因果关系的理解时，我们发现人的预期也会对因果关系的判断产生影响，这一点将在第 4 章进行探讨。这也是儿童和成人在因果关系判断过程中存在的另一个差异，因为他们对事件发生的可能性有着不同的预期。比如 5 岁的儿童在实验中会相信一个实际上不可能发生的事件是由魔法引起的，但 9 岁的儿童和成人则会意识到这不过是魔术而已。[17]

因果推理中的关联法本质上就是休谟提出的观点：如果人们经常看到一些事件同时发生，就会假设它们之间存在因果关系。[18]人类在做出这一假设时依据的案例数量比计算程序依据的数量少得多。但随着手中的数据越来越多，我们也会修正自己的观念，不过有时也会因为推理结论时太仓

促而找到错误的规律。当你穿了一双新球鞋并接连踢进了两个球时，你可能会觉得是新球鞋让你表现得如此出色，但在之后的十场球中你却一球未进，这时你可能就会重新思考之前在新鞋与踢球水平之间建立的联系。[19]

和感知能力一样，我们在很小的时候就已经具备从观察中推理原因的能力了。有一个实验专门测试了具备这种能力的最小年龄。有一个会播放音乐的盒子，如果将某个特定的木块放在盒子顶部，这个盒子就会播放音乐，但如果将其他木块放在盒子顶部则不会播放音乐。然后孩子们会看到每一个木块分别放在这个盒子上的结果，以及这些木块都放在盒子上的结果。实验发现，连两岁的孩子都能使用观察到的信息来判断出哪一个木块能够让盒子播放音乐。人们随后对 19 到 24 个月大的孩子[20]也进行了这一实验并得到了同样的结果。然后，又使用更简单的结构对 16 个月大的孩子进行了实验，实验结果表明，这些孩子普遍具有从变化的模式中推理原因的能力。[21]

但是，如果关联法是我们寻找原因的唯一方法，那我们要怎样区分事件中共同的原因（见图 2-1a）和共同的结果（见图 2-1b）呢？比如人们会因为失眠而去看电视、吃零食，而看电视、吃零食又会导致失眠。在现实生活中，即使我们所观察到的联系是一样的，我们也能够区分出不同的因果结构。如果我既喝咖啡又吃饼干，然后我发现大部分时候自己都精力充沛，而如果我只喝咖啡，大部分时候也会觉得自己精力充沛，那么我就可以推断出饼干并不影响我的精力是否充沛。

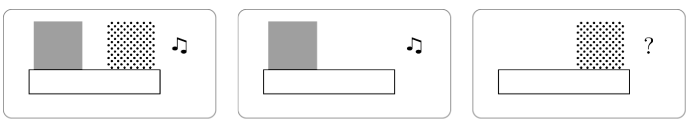


（a）共同的原因　　（b）共同的结果

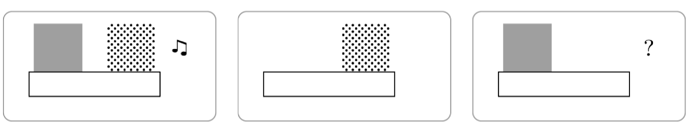
图 2-1　在这两个例子中，尽管因果结构发生了改变，但失眠都与另外两个活动相关联

这种推理方法被称为反向阻断法，也是研究人员在一项研究中证明过的方法，这项研究的参与者是 3 到 4 岁的儿童。[22] 这种推理方法的思路是，如果在两个因素都存在的情况下出现了某种结果，在只有第一个因素存在的情况下也出现了同样的结果，那么在没有见到阻断第二个因素所带来的影响的情况下，我们可以推理出第二个因素可能不是导致这种结果的原因。

我们再次使用了会播放音乐的盒子，在盒子顶部放上特定的木块，它就会播放音乐。孩子们首先看到木块 A 和木块 B 一起放在盒子上时，盒子播放了音乐，然后又看到盒子上只有木块 A 时也播放了音乐（见图 2-2a），这时孩子们就认为木块 B 不大可能是盒子播放音乐的原因。这个实验和前几个实验相比最主要的差别在于，在前几个实验中，孩子们看到了每一个木块单独放在盒子上的结果，也看到了这些木块共同放在盒子上的结果。而在这个实验中，木块 B 并没有单独放在盒子上，它只和木块 A 一起放在盒子上过，这就能让参与者根据木块 A 的效果间接判断出木块 B 的效果。但在这个实验中，3 岁儿童和 4 岁儿童之间还是有差别的。4 岁儿童认为“木块 B 也能让盒子播放音乐”的可能性要更小。



（a）反向阻断法



（b）间接推理法

图 2-2　参与者可以看到前两个实验的结果。在第三个实验中，他们必须预测如果将这个木块放在盒子上，盒子是否会播放音乐。实心图表示木块 A，点状图表示木块 B

在这个问题上，4 岁儿童的推理结果与成年人在研究中推理出的结果是一样的。[23]有趣的是，孩子们在推理因果关系时使用的也是间接证据。研究人员发现，即使孩子们看到两个木块都在盒子上时盒子播放了音乐，然后又看到盒子上放置其中一个木块后没有播放音乐（见图 2-2b），他们还是会推断那个没有单独放在盒子上的木块能够让盒子播放音乐。[24]

上述推理方法与关联法并不完全一致，因为我们可以从同一种关联关系中推理出多个结果。还有一种方法是因果模型法，它将因果推理和一个名叫“贝叶斯网络”的计算模型（将在第 6 章进行讨论）联系在了一起。[25]这种方法的理念是，可以把原因作为模型（这个模型可以向我们展示有多少件事物是相互联系的）的一部分，而不仅仅通过因素间的相关性或者各个因素间联系的强度来寻找事件发生的原因。图 2-1b 中的结构就是一个简单的例子。我们还可以扩展这个结构，增加导致失眠的原因（比如咖啡因和压力）和深夜吃零食的影响（比如发胖或长蛀牙）。这些结构可以帮助我们想出更好的干预措施，并帮助我们更好地利用这些干预措施来了解各个变量间的联系。

还有一种推理原因的方法是建立在作用机制上的。简单来说，就是原因是促成结果的一种途径，原因和结果是通过一系列能够导致结果发生的步骤连接在一起的。因此，如果跑步会让人心情变好，那么就一定存在一个跑步可以改变心情的过程，比如跑步能够释放体内的内啡肽。我们也许看不到这个过程中的每一个组成部分，但整个过程存在一个事件链，将原因和结果连接在了一起，原因通过这个事件链促成了结果的发生。[26]

然而，因果关系研究与协变关系研究所用的方法是不同的。在因果关系研究中，参与者需要向实验者提一些问题，以便能够解释某个事件是如何发生的。[27]在心理学文献中，这被称为因果推理。与之前的实验不同，接下来的实验任务要搞清楚的是“某个足球运动员为什么踢进了那一球”这种问题，而不是“一般情况下足球运动员进球靠的是什么”。以交通事

故为例，研究人员发现参与者的问题主要围绕有可能在事故中起作用的机制（比如驾驶员有身体缺陷吗），而不是倾向性问题（比如那条路上发生的交通事故多吗）。[28]在这种实验中，参与者必须去询问他们想要的信息。但在另一种实验中，我们为参与者提供了机制信息和协变信息，但参与者在确定原因的过程中利用机制信息的比重更大。

此外，我们还把观察到的信息和已知的信息进行了融合，并且掌握了一些与相关性和作用机制有关的知识。因此，我们不会仅依赖某一种类型的证据来进行研究。事实上，我们还做了很多其他研究工作来考察各种信息是如何（而非是否）结合在一起的。比如说，有一些实验表明，参与者对"原因和结果之间是否存在关联机制"的态度会影响他们对事件之间关联性的看法，但如果两个事件的关联性很弱，那么参与者将不会受到影响。[29]实际上，参与者在对观察序列（比如某些常见的/不常见的症状）进行评估时，可能会把事件之间已知的关联以及存在某种关联的可能性也考虑进去。[30]

然而，人们在如何得出各种关联关系（后面将其统称为模型或因果结构）的问题上却产生了分歧。有一种观点是，我们应该先收集数据，然后根据这些数据来选择最可能出现的结构，或者与我们的观察结果最一致的结构。如果我们知道狗听到大的声响就会叫，也知道摔门会发出大的声响，那我们就能缩小范围，缩减这些事件相关联的可能性方式，并且排除掉"狗引起了声响"这样的模型。[31]还有一种观点认为，很多时候我们都是依靠各种假设来思考问题的，所以我们应该先提出一个有可能出现的结构，然后再根据了解到的新信息不断改变这个结构。[32]

这些实验大多很简单，我们能够通过可控的实验环境来分离不同变量对结果的影响。但在现实中，我们很少专门研究一个（已经被确认为潜在原因的）事物对另一个（已经被确认为潜在结果的）事物的影响程度。比如某个人突然开始头疼，这时候他就必须回想各种可能引起头疼的因

素。同样，要想找出某种药物的过敏反应，就要从众多服用该药物的人中找到他们服药后不断出现的共同症状。因果推理活动通常包含两部分：寻找结构和寻找影响力。结构会告诉我们什么原因导致了什么结果，而影响力会告诉我们这个原因在多大程度上导致了这个结果（比如某种药物产生某个副作用的频率是什么；或者在收益报告发布之后，某个股票的价格会上涨多少）。寻找结构与影响力的过程是无法完全分离的，因为影响大的原因比影响小的原因更容易找到。很多心理学实验都会重点研究参与者对影响力的评估，这可能也是人们关注协变关系而不是因果机制的原因。

比如说，你注意到自己在跑步时经常打喷嚏。如果无法改变跑步的环境（比如室内或室外、春天或冬天，等等），你就无法发现你打喷嚏的原因其实是季节性过敏，而不是对运动的反应。在那些简单的案例中，孩子们能够仅通过观察一系列事件就推理出正确的因果结构。但如果仅依赖对数据的观察，就会让人们混淆因果关系，比如仅因为两个结果有一个共同的原因并且经常同时出现，就错误地以为这两个结果互为对方产生的原因。

原因之所以如此重要，关键理由之一是我们可以利用它来设计出有效的干预措施，从而控制我们周围的世界。但是，干预措施也能反过来帮助我们找到事件发生的原因。在前面的心理学实验中，我们巧妙地将这个世界划分成了各种可能的原因和结果。当不知道哪个是原因哪个是结果时，我们控制这些因素，并测试当出现或缺少不同的因素时会出现什么样的结果，这样就能区分出那些看似相同的因果结构。一些研究发现，如果我们允许参与者去干预一个体系的运转过程而不仅仅是让他们观察，那么他们推理因果关系的准确度就会提高。[33]

有一项研究用一个简单的齿轮玩具验证了上述结论。因为这个玩具有两个齿轮和一个开关，所以它能够实现多种因果结构：一个齿轮让另一个齿轮转动、开关让两个齿轮分别转动、开关让两个齿轮一起转动。学龄

前儿童仅通过观察他人对这个玩具的干预就能了解这些更为复杂的因果结构。[34]然而，看和做（观察和干预）是有差别的，自己实施干预和看别人实施干预也是有差别的。当你自己选择并实施了一项干预措施时，你可以提出假设并进行验证，并且能够控制那些你认为可能影响结果的因素。事实上，在有些实验中，无论是儿童还是成人参与者都能从自己的干预活动（而不是他人实施的干预活动）中更好地把握事物之间的因果结构。[35]

2.2 责任的划分

假如你有一台非常考究的咖啡机，这台机器的热度达到萃取咖啡的要求之后只能持续很短的时间，你必须在机器过热之前把咖啡萃取出来。你的朋友在萃取了一杯咖啡后没有关机器，你去萃取咖啡的时候机器已经过热，所以那天早上你没喝成咖啡。那么，是谁造成你那天没喝成咖啡的？是因为你的朋友没有早点关掉咖啡机，还是因为厂家生产了一台有缺陷、不能重度使用的机器？

这就是因果关系中的归因问题：要确定是谁或者是什么导致了特定的事件。也就是说，我们想知道的不是一般情况下咖啡机出故障的原因，而是这个案例中的咖啡机为什么会出故障。这与我们分析交通事故中的责任，或者分析某人为什么开会迟到是同一种类型的推理活动。这种类型的因果关系称为实证式因果关系（as token causality），与类级别因果关系（type-level causality）刚好相反。类级别因果关系指的是一般情况下会出现的情况。比如，因分心驾驶而导致的交通事故，与 Susie 在开车时发短信结果撞上了 Billy 的车是不同的。我们将在第 8 章深入讨论实证式因果关系。

不过，在进行责任划分时，我们不能单单列出可能引发事故的各种原因，还要考虑道德因素或过失程度。此外，有些事件中可能存在因果关系，但这里面的原因并没有实际的责任。比如说，你导致了一场交通事故

却不用承担责任，因为你当时及时踩了刹车，但因为刹车失灵而撞上了另一辆车，这时你是没有责任的（我们将在第 8 章解释为什么这可能是汽车生产商的责任）。关于责任划分和归因问题的研究大多是在哲学领域进行的，但这些研究通常基于直觉或者人们“应该会”有的想法，而不是通过实验收集的数据。比如我们所说的“笔的问题”。某大学哲学系有一名接待员，她的办公桌里装满了笔。行政助理们需要用笔的时候直接从接待员那里拿就可以了，但是教授们则需要自己买笔。但实际上，教授和助理都会从接待员那里拿笔。有一天，一名教授和一名行政助理拿走了接待员手中最后的两支笔。然后，接待员接到了一通重要的电话，结果却没有笔来记电话中的内容。这个情况是谁造成的呢？[36]

关于这个问题，每个人的直觉可能不一样。我们并不清楚人们对这个问题的主导性看法是什么，也不知道这个问题是否有正确答案。研究这些问题的哲学家常常假定人们对这些问题有一个共同的直觉。心理学家则通常通过实验来验证这些观点，但是大多数实验的参与者都是大学生，所以我们无法确定能否由此推断出整个人群的道德考量（可能大学生对行政助理和大学教授的道德问题有强烈的先入为主的情感）。人们用实验法来回答哲学问题的做法已经越来越常见，并且时常用实验法去验证一些通常被认为理所当然的直觉，这就催生了哲学上的一个分支——实验哲学。实验哲学的主要研究领域之一正是这类道德评判，而这也是哲学和心理学研究的交叉领域。

有一个重要发现叫作“副作用效应”（也叫“诺布效应”）[37]，它的主要内容是，如果某个人的行为无意间带来了积极的效应，人们不会将这种效应归功于这些行为；但如果这个人的行为无意间带来了消极效应，那么人们就会认为这些行为是有意的，并且将责任归咎于这些行为的发出者。有一项实验告诉参与者：公司的 CEO 并不关心他们关于提高利润的最新提案是否对环境有利，他们只关心利润。结果当环境遭到破坏时，参与者

往往会责怪这位 CEO；但是当环境得到改善时，参与者却并没有表扬这位 CEO。人们又做了类似的实验并得到了同样的结果：无意间带来积极效应的行为不会得到表扬，但无意间带来消极效应的行为却会受到批评。[38]心理学家的实验表明，与无意的行为相比，有意的行为更容易成为事件的原因和责任主体。[39]这项研究很有名，因为实验的参与者不是大学生，而是纽约一个公园里的人，但研究人员并没有告诉我们这些参与者的所属区域及人口特征方面的信息。[40]

我们要了解的关于“动机”的第二个方面是，人们想要得到的结果和实际产生的结果之间的差别。正如驾驶员试图把车停下来却因为机械故障而没能停下一样，有时候人的动机可能是好的，但结果却是坏的。如果某个人的动机是好的，但他的行为却带来了不好的结果，那么他是否应该像那些有意造成不良后果的人一样受到责备呢？针对这类问题的一些研究表明，动机与结果之间的相互作用比道德评价与结果之间的相互作用更能解释人们的评判。举个例子，某人有意造成某种伤害但没有成功，却有人因为其他原因而受到了伤害。与没有任何人受到伤害的情况相比，在这种情况下，人们划分给有意造成伤害的人的责任要少一些。[41]考虑到这一结果，我们似乎可以理解为什么作弊失败的人比作弊成功的人受到的责备要少一些——尽管试图作弊的人都会受到责备。

关于副作用效应的解释有两种：一种是它取决于人们的行为是否是有意的，另一种是人们的行为是否违反了社会规范。[42]如果你的行为符合社会规范（考试不作弊、不乱扔垃圾，等等），那么你不会因为自己的行为而受到褒奖，因为这是正常行为。然而，如果你为了走捷径而踩了一些花草，就会受到责备，因为你的行为违反了社会行为标准。有一个违背了社会规范却没有造成任何后果的例子——在没有任何车辆行驶的、空旷的柏林街头，在没有斑马线的地方横穿马路（这种行为在柏林是不允许的）。在这个例子中，没有任何造成伤害的动机，也没有人受到任何伤害，但是它

依然违背了社会规范。我们通常不会因为没有发生的事而去问责，但是这种行为可能会受到责备，因为这样有可能会造成伤害。这也许就是不遵守交通规则的人会被别人责骂的原因。

有一个实验明确地验证了行为规范、对行为的道德评判和行为结果之间的关系。[43] 在这个实验中，有一组学生拿到了期末试卷。我们可以根据同一问题的不同答案而得到不同的情形。首先，大部分学生可以选择作弊或者不作弊。然后，一个名叫 John Granger 的学生可以选择随大流（大部分人作弊时他也作弊，大部分人不作弊时他也不作弊）或者不随大流（大部分人作弊时他不作弊，大部分人不作弊时他作弊）。在他的考试成绩和评分机制的双重作用下，期末考试成绩仅次于他的那个同学，因为一线之差而没有达到医学院要求的最低平均绩点（GPA）。那么问题来了，如果 John Granger 需要承担责任，那么他是在什么情况下造成了这一结果的？有趣的是，规范性并没有对因果关系或责任划分产生过多的影响。相反，参与者的判定主要依据他们对 Granger 行为的评估，如果他们认为 Granger 的行为很坏，那么这种行为就更能引起这样的结果，也更应该加以谴责。然而，当大部分参与者都作弊而 Granger 没有作弊时，人们就会认为他的行为不至于受到谴责。

有证据表明，影响责任判定的因素有很多，比如规范、动机和结果等，然而做出这些判定的过程还在研究当中。尽管近期的研究将责任判定当作了一种包含多个步骤和流程的社会行为，但是绝大部分实验研究还是主要关注结果并致力于理解各种直觉。[44]

2.3　文化

当有些研究指出“90%的参与者都认为是那个司机引起了这场交通事故”时，这里的参与者指的是什么人？心理学研究的绝大部分参与者都是

西方大学生。[45]这并不奇怪，因为这个领域的大部分研究工作都是在高校开展的，我们通常都能找到足够的学生参与者。某些情况下可能会存在一种普遍的现象，但这并不意味着每个人所感知的、所判定的因果关系都是一样的，更不要说那些 18 岁到 21 岁的大学生了。这就限制了我们所讨论的那些研究成果的普遍适用性。为了了解这个问题的影响，有些研究人员对比了不同文化背景的参与者对因果关系的感知和判定。

一个重要的文化差异是，参与者认为哪些因素与结果存在因果相关性。[46]如果某个游泳运动员赢得了一次奥运比赛，人们可能会说她之所以能获胜是因为参赛运动员的总体实力比较弱，或者是因为她有家人的支持（环境因素），或者是因为她有游泳天赋（个人禀赋）。这些因素可能都为她的成功做出了贡献，但是差别在于哪些因素是最重要的。为了验证这一点，Michael W. Morris 和 Kaiping Peng（1994）分析了汉语报纸和英文报纸上关于同样的一些刑事案件的报道，他们发现英文报纸上提到性格因素（比如凶手十分愤怒）的比例要比中文报纸高得多，而中文报纸则往往强调环境因素（比如凶手刚刚失业）。Michael W. Morris 和 Kaiping Peng 还让中国学生和美国学生对各种影响因素的重要性进行了评分，结果与上述研究一致。在其他针对东西方文化的对比研究中，人们也发现了同样的现象。[47]

然而，这些文化差异似乎还会随着年龄的增长而发生变化。Joan Miller（1984）是第一批研究这一现象的研究者之一，她对比了四个不同年龄段（8 岁、11 岁、15 岁及成年人）的印度参与者和美国参与者，发现在 8 岁和 11 岁这两个群体中，印度参与者和美国参与者几乎没有差异。研究者让这些参与者解释一下他们所认识的某个人为什么会做好事，另一个人又为什么会做坏事，年龄越大的美国参与者越强调个性特征的作用（比如这个朋友心地比较善良），而年龄越大的印度参与者则越强调环境的作用（比如他刚刚换了工作），其中成年美国参与者和成年印度参与者的

差别最大。这可能是因为人们的观点真的随着年龄的增长发生了变化，也可能是因为随着年龄的增长，人们越来越明白自己该说什么，不该说什么。我们知道仅仅参与一项研究也能影响人们的行为，因为参与者可能会尽量按照他们所理解的研究者的信念做事（就是尽量让研究者高兴），也可能会与研究者对着干。在一项研究中，研究者仅仅改变了问卷的抬头就导致参与者的答案重心发生了变化。[48]

在归因问题上，社会暗示似乎会对人们所强调的因素（比如新闻中报道的内容）产生一定的影响，并且对人们如何描述这些因素的重要性（环境和个性的影响程度有多大）也有一定的影响。但是，这一现象背后的作用机制却无人知晓。不久前，一些证据表明人们对舆论的认识（你认为你所处的社会群体会持什么样的观点）会对文化差异产生影响。[49]也就是说，尽管这些研究的结果与早期 Morris 和 Peng 的发现是一致的，但是这些参与者实际上可能持有同样的观点，只不过他们所认为的全体中国人的观点和他们所认为的全体美国人的观点不同而已，而这种对群体所持观点的认识可能解释了他们为什么会做出不同的判断。

现在我们似乎明确了一点：人们对交通事故（一个涉及很多社会与文化因素的事件）中的过错方可能会有不同的判断。反对分心驾驶的人可能会抓住驾驶员开车时发短信这个事实不放，另一个人可能会因为汽车制动系统故障而认为是汽车制造商的过错。有人假设，个人主义文化和集体主义文化的差异是导致归因差异的根源，所以这种差异只会体现在我们对社会性事件（动物群体或人类群体之间的互动）的感知当中，而不会体现在对物理性事件（物体的移动）的感知当中。我们对物理事件的感知似乎不会受到文化差异的影响，但最近有研究发现，人们在感知物理事件时，眼球的运动存在文化上的差异（不同文化的人可能会将注意力放在同一场景的不同位置上）。[50]

2.4 人的局限性

尽管我们的一个主要目标（也是长远目标）是设计出能够复制人类思维的算法，但是人类思维在很多方面都不如计算机程序，因为计算机程序的运算行为是完全可控的，并且能够完全依照制定好的规则来运行。虽然我们能够从很少的观察数据中快速找到因果关系，但是我们所找到的因果关系并不总是正确的。更令人苦恼的是，我们经常会犯同样的错误，即便我们已经意识到了这一点。第 3 章会讲到，很多认知偏见会导致我们看到一些并不存在的相关性，因为我们经常会寻找一些信息来证实自己的信念（比如找一些同样觉得针灸有效果的人），或者更重视那些能够证实我们信念的信息（比如在收银台排队时，我们只会注意到比自己的队伍结账速度快的队伍）。有些因素让人们很难把握事件发生的原因，比如原因和结果之间存在很长的延迟，或者因果结构很复杂。这些因素要求人们解开很多复杂的关联关系，同时还可能会让事件之间的联系变得模糊。但是，即便是面对一个原因和结果之间没有延迟的、简单的因果结构，我们仍有可能在因果推理过程中被某些谬论误导。

祸不单行是真的吗？打碎一面镜子就会带来七年的霉运吗？口香糖吞进肚里真的需要几年才能消化吗？在错误的因果观念中，最有说服力的形式之一就是迷信。一定没有人统计过一个人在打碎一面镜子之前或之后的七年中遇到的倒霉事，也没有人将打碎镜子的人群和没有打碎镜子的人群在七年中遇到的倒霉事做过比较。既然如此，为什么那么多通常都很理性的人还会继续相信这种说法呢？

有些类似于这样的迷信可以用某种认知偏见来解释，从而让我们能够识别出本来没有联系的事件之间被人为杜撰出来的各种错误的相关性。也就是说，人们之所以会在打碎镜子之后注意到更多倒霉的事情，是因为人们对这些事情的警觉性提高了。更糟糕的是，如果你相信之后会有七年

的霉运，那么你可能会把一般情况下不会注意到的或者不认为是倒霉的事情也当作霉运。但还有一些情况下，迷信思想会产生一种安慰剂效应。

在医学上，仅仅是“接受治疗”这一行为就可能对病人产生影响，所以在研究药效时必须找一个参照物。我们将某种药物的疗效与另一种类似的、已知没有效果的治疗方法相比较。[51] 比如说，我们可以对比阿司匹林和糖丸治疗头痛的效果，而不是将阿司匹林的疗效与不采用任何治疗方法的情况相比较，因为这样我们才能控制服药这一行为本身对病人的影响。这就是在没有与任何治疗方法做对比的情况下，“实验疗法让病人的症状减轻了 10%”这种结论没有任何意义的原因。有时候，病人已经知道了他们正在服用安慰剂，即使这个安慰剂对他们的病症没有任何帮助，他们依然会发现安慰剂效应。[52]

同样，当你相信自己的幸运之笔，或者篮球赛前的一些仪式能够帮助你得分时，它们就有可能真的带来这种效应。值得注意的是，给你带来好结果的并不是这个物件或仪式本身，而是你认为这个物件或仪式会起作用的信念。这种信念会给你带来某些感觉，比如减轻压力或者胜券在握的感觉，而这些感觉又给你带来了好的结果。[53]

你可能会觉得以上这些听起来都有点道理，但是“7”这个数字对你来说意义特殊，而且这不是巧合。你听到的所有好消息都正好是分钟的个位数是 7 的时候，出现这种情况的概率有多大呢？一旦你产生了这样的迷信，就会特别关注与这个迷信一致的事情，也更容易记住这样的事情。也就是说，你会开始忽略与这个迷信不一致的事情（比如与 7 无关的好事）。人们倾向于去寻找并记住一些证据，以此来支撑自己的信念，这种倾向叫作证实性偏见，我们将在下一章进行更全面的讨论。这种偏见可能会导致人们产生一些无害的错误观念，但也有可能会强化一些有害的偏见。

这有点像是成见威胁现象：当一个人知道自己属于一个具有负面特征的群体时，他就会害怕去证实那些成见。在一次实验中，实验者让参与

者做一份数学试卷。在考试前，实验者告诉一部分女性参与者，男生和女生做这份试卷所考出的成绩是不一样的（有趣的是，实验者并没有告诉她们到底是男生考得好还是女生考得好）。[54] 然后又对另一部分女性参与者说，男生和女生做这份试卷所考出的成绩没有差别。结果，被告知男生和女生的成绩没有差别的女性参与者考出的成绩和男生平分秋色，而被告知男生和女生考出的成绩有差别的女性参与者考出的成绩要比男生差得多。这种错误的因果信念可能会带来实际的影响。我们在之后的章节中将会看到，建立在错误的因果关系之上的政策最多不过是无效政策，而使用错误的因果关系则可能会造成冤案，正如我们在第 1 章看到的那样。

一种无代价的或者不显眼的仪式可能无伤大雅（手指交叉祈祷似乎并没有多大的害处），但这种行为最终会导致人们去依赖一种微弱的联系，并有可能高估自己的力量（一个人控制或预测事件的能力）。[55] 我们会提出假设并寻找迹象来证实这些怀疑，但是严谨的因果思维需要我们认识到这种行为可能会让我们陷入偏见，并且必须接受与我们的信念相反的证据，后面将会介绍如何做到这一点。

注释

1. Caporael（1976）。
2. Matossian（1989）。
3. Spanos 和 Gottlieb（1976）。
4. Spanos 和 Gottlieb（1976）；Woolf（2000）。
5. Sullivan（1982）。
6. Schlottmann 和 Shanks（1992）。
7. Roser 等（2005）。
8. Michotte（1946）。
9. Leslie（1982）；Leslie 和 Keeble（1987）。值得注意的是，其他人在 6 个月的婴儿身上做的研究也得出了类似的结论，他们不仅发现了启动序列，还发现了“追逐”序列（Schlottmann 等，2012）。
10. Oakes（1994）。

11. Cohen 等（1999）。
12. Schlottmann 等（2002）。
13. Schlottmann（1999）。
14. Badler 等（2010）。
15. Badler 等（2012）。
16. 关于机械理论和共变理论之间的联系，参见 Danks（2005）。
17. 有趣的是，尽管 6 岁的孩子一开始表现出了对魔术的怀疑，但在见到表面上相反的证据后，他们却愿意改变自己的信念（Subbotsky，2004）。
18. Rescorla 和 Wagner（1972）；Shanks（1995）。
19. 想要了解这些心理学理论的详细信息，参见 Cheng 和 Novick（1990，1992）（概率差异），Cheng（1997）（因果力），Novick 和 Cheng（2004）（因果力）。
20. Gopnik 等（2001）；Sobel 和 Kirkham（2006）。
21. Gweon 和 Schulz（2011）。
22. Sobel 等（2004）。
23. Shanks（1985）；Spellman（1996）。
24. Sobel 等（2004）。值得注意的是，在训练阶段，实验人员给孩子们看了一个单独分开的木块，而且孩子们看到单独使用这个木块就可以开动机器。所以，孩子们知道一个木块是有可能单独起作用的。
25. 想要回顾相关信息，参见 Holyoak 和 Cheng（2011）。
26. Ahn 和 Kalish（2000）。
27. Ahn 和 Bailenson（1996）。
28. Ahn 等（1995）。
29. Fugelsang 和 Thompson（2003）。
30. Griffiths 等（2011）。想要了解机械信息和共变信息是如何结合在一起的，参见 Perales 等（2010）。
31. 参见 Gopnik 等（2004）；Griffiths 和 Tenenbaum（2005）。
32. 想要了解这方面的信息，参见 Lagnado 等（2007）。
33. Lagnado 和 Sloman（2004）；Steyvers 等（2003）。
34. Schulz 等（2007）。还有其他研究曾将干预措施的作用和贝叶斯网络的形式主义联系在一起。参见 Gopnik 等（2004）；Waldmann 和 Hagmayer（2005）。
35. Kushnir 和 Gopnik（2005）；Sobel 和 Kushnir（2006）。
36. 想要获取免费用笔问题的完整文本，参见 Knobe 和 Fraser（2008）。
37. 想要了解诺布效应最初的发现以及这项研究的详细信息，参见 Knobe（2003）。

38. 具体案例，参见 Knobe 和 Mendlow（2004）；Nadelhoffer（2004）；Uttich 和 Lombrozo（2010）。
39. Lagnado 和 Channon（2008）。
40. 在原文中，作者并没有说明调查中提到的是哪一个公园，也没有说明被调查对象的年龄和人口统计学信息。后来的研究（Meeks，2004）指出，调查中提到的公园是位于纽约大学中央的华盛顿广场公园和同样吸引了很多学生和年轻人的汤普金斯广场公园。在一次访谈中，Knobe 提到他的调查对象既有来自于中央公园的参与者也有来自于华盛顿广场公园的参与者，还提到他的调查结果发现来自这两个公园的参与者的回答存在差异，而且这一差异在统计学上具有显著意义。当然，这些都没有写进他的文章中。
41. Cushman（2008）。
42. 想要了解更多关于规范的观点，参见 Hitchcock 和 Knobe（2009）。
43. Alicke 等（2011）。
44. 想要了解更多信息，参见 Malle 等（2014），也可参考与该杂志同期发表的针对该文章的诸多评论文章。
45. Henrich 等（2010）。
46. Choi 等（2003）。
47. Choi 等（1999）；Morris 和 Peng（1994）。
48. Norenzayan 和 Schwarz（1999）。
49. Zou 等（2009）。
50. 虽然有些研究已经显示不同的人在归因的过程中会存在一些文化上的差异，这些差异主要表现在对原因的解释（Peng 和 Knowles，2003）以及人们在观看一个场景时的眼球运动上（Chua 等，2005），但大部分研究尚未证实人们在对物理事件的归因过程或者在感知原因的过程中存在文化上的差异。
51. 安慰剂的构成成分并不总是那么容易回答的。在一种情况下是安慰剂的事物在另一种情况下也许就不是了。想要了解更多信息，参见 Grünbaum（1981）和 Howick（2011）。
52. Kaptchuk 等（2010）。
53. Damisch 等（2010）。
54. Spencer 等（1999）。
55. Pronin 等（2006）。

第 3 章　相关性

为什么有那么多因果关系被搞错？

2009 年，研究人员发现一种叫 XMRV 的病毒与慢性疲劳综合征[1]（CFS）有着惊人的联系。尽管美国有数百万人患有这种疾病（其特征是长时间的严重疲劳），却没有人知道病因。由于病因不明，所以这种疾病的预防和治疗工作都遇到了阻碍。这种疾病到底是由什么引起的呢？人们提出了很多假设，其中包括病毒、免疫缺陷、基因和压力等。[2]然而，这种疾病不仅病因不明，就连诊断也十分艰难，因为无法通过检测某个生物指标来确诊。还有很多病例都没有被发现，而且有可能 CFS 其实只是很多疾病的一个总称。[3]

在这种情况下，Judy Mikovits 带领的研究团队的发现引起了人们的注意。他们发现，在 101 个慢性疲劳综合征患者当中，有 67%的人身上带有 XMRV 病毒；而在由 218 人组成的控制组中，只有 3.7%的人身上带有 XMRV 病毒。尽管 XMRV 病毒论并不能解释所有的病例，但可能有一部分病人的慢性疲劳综合征就是由 XMRV 病毒引起的，而且控制组中那些带有 XMRV 病毒的人也可能是没有诊断出来的慢性疲劳综合征患者。这些数据对于一个如此难以解释的疾病来说十分重要，并且催生了一大批想要证实这些结论的研究活动。很多研究都没能找到慢性疲劳综合征和 XMRV 病毒之间的联系，[4]但研究人员在 2010 年发现了一种类似的病毒，

这种病毒在慢性疲劳综合征患者身上出现的比例（86.5%，37 人中有 32 人带有这种病毒）明显高于在健康的献血者身上出现的比例（6.8%，44 人中有 3 人带有这种病毒）。[5] 这一发现推动了新一轮的假设并催生了更多的研究活动，研究者们都想要证实或推翻这种病毒和慢性疲劳综合征之间的联系。

人们假定这种极强的相关性意味着 XMRV 病毒就是慢性疲劳综合征的病因，所以针对这种病毒的治疗方案也许能够最终治愈慢性疲劳综合征。一些患者十分渴望找到一种治疗方案来治愈这种令人虚弱的不治之症，他们甚至基于对 XMRV 病毒的研究向医生索要抗逆转录病毒药物。很多慢性疲劳综合征患者的血样中都有一种相同的病毒，这一发现十分有趣并且值得进一步研究，但是我们无法仅利用这一相关性来证明这个病毒就是罪魁祸首，也无法证明抗逆转录病毒药物就是一种有效的治疗方案。也有可能是慢性疲劳综合征导致免疫系统受损，从而让人们更容易感染这些病毒。即便病毒和疾病之间存在某种因果关系，这种很强的相关性也并不能告诉我们前因后果——这个病毒到底是因还是果，还是说二者是由同一个原因导致的共同结果。

2011 年，关于慢性疲劳综合征和某种病毒之间相关性的两项研究，在经历了很多争议和公开的辩论之后都被撤回了。对于 Mikovits 医生的研究，一开始是部分撤回，但最终还是由杂志社在未经作者同意的情况下全部撤回了。[6]事情是这样的，Mikovits 医生研究的血样遭到了 XMRV 病毒的污染，导致两组样本出现了表面上的差异。[7] 除了样本污染问题以外，人们还怀疑可能存在伪造数据问题，因为有一个图例省略掉了关于如何准备样本的信息，并且有人指出同一图例在不同的地方被贴上了不同的标签。[8]此外，2012 年有一项研究邀请了多个团队（包括 Mikovits 的团队），各个团队所用的分析样本都是盲样，研究结果表明慢性疲劳综合征和 XMRV 病毒之间没有任何联系。[9]

最初的那个发现引起了人们的极大关注，然后就极具戏剧性地公开上演了各种分歧。这一切都向我们展示了看似极强的相关性能够带来的巨大影响。

“相关性不是因果关系”，统计学专业的学生对这句话已经烂熟于心，但即使是那些理解并且赞成这个说法的人有时候也会忍不住把相关性当成因果关系。研究人员在报告相关性时常常会附上很多说明，以此来解释为什么这些相关性不是因果关系，以及还缺少什么信息。但这些相关性仍然会被人们解读成因果关系，并被当作因果关系来使用（有时一篇科技论文和大众媒体对这篇论文的报道之间都存在着巨大的差异）。极强的相关性可能很有说服力，也许还可以让我们做出一些成功的预测（尽管慢性疲劳综合征的案例并不是这样的），但它无法告知我们事物的工作原理，也无法告知我们如何采取干预措施来改变这个事物的运行机制。慢性疲劳综合征和 XMRV 病毒之间的表面联系并不能说明我们能够用治疗这个病毒的方法来治好慢性疲劳综合征，但病人却认为这是可行的。

表面上的相关性也许可以用无法测定的原因来解释（如果省略关于吸烟的数据，就会导致很多其他因素和癌症之间出现相关性），但是两个本没有任何关系的变量之间也可能会出现一些虚假的联系。相关性可能是巧合（一周遇到某个朋友好几次），也可能是通过研究方法而人为导致的（调查问卷中的某些答案选项可能具有偏向性），还可能是由于失误和操作不当导致的（计算机程序中的漏洞）。

即便如此，相关性仍然是我们所能得到的最根本的发现之一，也是能证明因果关系的一个证据。在这一章，我们将探讨相关性的定义和用途，以及一些出现相关性但背后却没有因果关系的情况。

3.1 相关性是什么

X 和癌症有关，Y 和中风有关，Z 和心脏病发作有关。这三句话描述了三个相关性，告诉我们两个现象是相关的，却没说它们是如何关联在一起的。

两个变量相关的基本意思是，一个变量发生的变化与另一个变量发生的变化是有关联的。比如说，孩子们的身高和年龄相关，因为随着年龄的增长，孩子们的身高也会增长，这样他们才能慢慢长大。这些相关性可能存在于不同的样本之间（一次测量多个不同年龄的孩子），也可能存在于同一样本的不同时间段之间（在同一个孩子的不同年龄段多次测量他的身高），还可能存在于不同样本的不同时间段之间（在多个孩子的不同年龄段多次测量他们的身高）。然而，身高和出生月份之间却没有长期的相关性。也就是说，即使我们改变了出生月份，我们的身高也并不会发生有规律的变化。图 3-1a 展示了年龄的变化是如何与身高的变化相对应的。随着一个变量的上升，另一个变量也会上升。图 3-1b 展示了身高和出生月份之间的关系，这幅图看起来就像是一堆随机放置的黑点，而且身高并没有随着出生月份的变化而发生相应的变化。

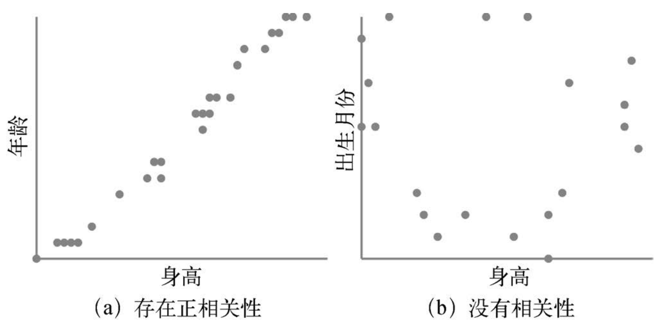


图 3-1　年龄和身高是相关的，但身高和出生月份是不相关的

这意味着如果知道一个孩子的年龄，我们就能大致预测出他的身高，但如果我们只知道他的出生月份，则无法预测出他的身高。上面那些黑点的排列越接近一条直线，我们的预测就会越准确（因为排列越接近直线说明二者之间的相关性越强）。相关性的主要用途之一就是预测，而且有时可以在没有因果关系的情况下做出预测。当然，这些预测并不总是成功的。

相关性很强时看起来也会很明显（如图 3-1a 所示），但我们也需要用一些方法来测量相关性的强度，以便对其进行定量比较和评估。表示相关性的指标有很多，但是最常用的指标之一就是皮尔逊相关系数（通常用字母 r 表示）。[10] 这个系数的数值介于 1 和–1 之间，系数 1 表示变量之间存在完美的正相关性（一个变量发生正向变化会直接引起另一个变量发生相应的正向变化），而系数–1 则表示变量之间存在完美的负相关性（如果一个变量减小，则另一个变量一定会增大）。

简单来说，皮尔逊相关系数是指两个变量如何通过各自的变化而发生共同的变化（这两个数值称为协方差和方差）。比如说，我们可以记录一组学生的学习时间和期末考试成绩，以便了解二者之间的关系。如果我们只有一组考试成绩的数据和一组学习时间的数据，而没有将相应的考试成绩和学习时间一一对应，那就无法确定二者之间是否具有相关性。这是因为我们只能看到个体在每一个变量上的变化，而没有看到这两个变量是如何共同发生变化的。也就是说，我们无法得知更长的学习时间是否对应更高的考试成绩。

3.1.1　没有变化就没有相关性

比如说，你想知道如何写申请才能获批某项资助，所以就去找所有申请到这项资助的朋友，询问他们自认为让他们成功获批的因素。这些朋友在资助申请中都用了 Times New Roman 字体，其中有一半人说每页至少要有一个图表，还有三分之一的人建议你在截止日期的前一天提交申请。

这是否意味着在这些因素和资助成功获批之间存在相关性呢？不是的，因为结果没有发生变化，所以我们无法确定是否还有其他因素和结果有关。如果我们观察到在连续几个气温为华氏 80 度的日子里，街道的某个拐角处都正好有两个冰淇淋小贩，我们不能由此就对天气和冰淇淋小贩之间的相关性发表任何看法，因为这两个变量的数值（气温或者冰淇淋小贩的数量）都没有发生过变化。同样，如果我们只看到一个变量发生了变化（比如冰淇淋小贩的数量总是两个，但气温却在华氏 80 度到 90 度之间发生变化），也不能得出任何结论。图 3-2 所示的正是这种情况，没有发生变化的数据在图中是一个不变的黑点，而只有一个变量发生变化则呈现为一条横线。[11] 之前那个申请资助的例子也是如此。因为所有的结果都是一样的，所以我们无法预测如果改变字体会出现什么情况，也无法预测如果正好在截止时间之前的那一刻提交申请书会出现什么情况。



（a）没有变化　　（b）一个变量发生了变化

图 3-2　如果两个变量没有共同发生改变，我们就无法找到它们之间的相关性

然而，大部分人都只关心导致某个结果的原因是什么。比如人们常常会问那些成功人士是如何取得成功的，然后试图通过复制他们的做法来取得成功。这一做法从很多方面来讲都是存在严重问题的，比如人们不善于分辨哪些因素是重要的，哪些因素是不重要的，并且往往会低估机遇的

重要性而高估自己的技能。[12] 于是，我们会将那些仅仅与我们所关心的结果同时出现的因素误认为是导致这种结果的因素，而且还会发现一些并不存在的表面联系。

有人曾经问过：人们在其他领域所取得的专业性成就与他们所接受的音乐教育之间是否存在相关性？即便很多成功人士（无论我们如何定义成功）也会演奏乐器，我们也不能说这二者之间就存在相关性——更不要说因果关系了。如果我们直接去访问一些成功人士，问问他们是否认为音乐有助于提升他们的其他能力，那么一定有很多人能从这两件事之间归纳出一些联系。但是，如果我们问他们是否认为下棋、跑步或者喝咖啡有助于提高他们的其他能力，他们也完全能找到一些联系。

对这本书来说最关键的是，我们不能仅仅调查那些成功者所谓的秘诀，因为有些人可能做了完全相同的事情却没有获得成功。也许所有申请资助的人都用了 Times New Roman 字体（所以如果我们去询问那些没有成功的人，他们会建议我们使用其他字体），也许这些成功者使用了过多的图形但依然得到了资助。如果不能全面分析成功的例子和失败的例子，我们甚至都不能确定事物之间是否存在相关性。

3.1.2　相关性的测量与解释

比如我们调查了一些学生在期末考试前喝了多少杯咖啡，然后又记录了他们的期末考试成绩。这个案例的假设数据如图 3-3a 所示，两个变量之间的相关性非常高，相关系数接近 1（确切地说是 0.963），所以图上的黑点似乎紧紧地聚在一条无形的直线两侧。如果我们将这一关系反过来（于是不喝咖啡的学生考试成绩成了 92 分，而喝 10 杯咖啡的学生考试成绩则为 10 分），建立一种负向联系，那么相关变量的变化幅度是一样的，唯一改变的就是相关系数的符号。在这种情况下，这个相关系数会接近 −1（−0.963），刚好是正相关数据图水平翻转过来的样子（如图 3-3b 所示）。

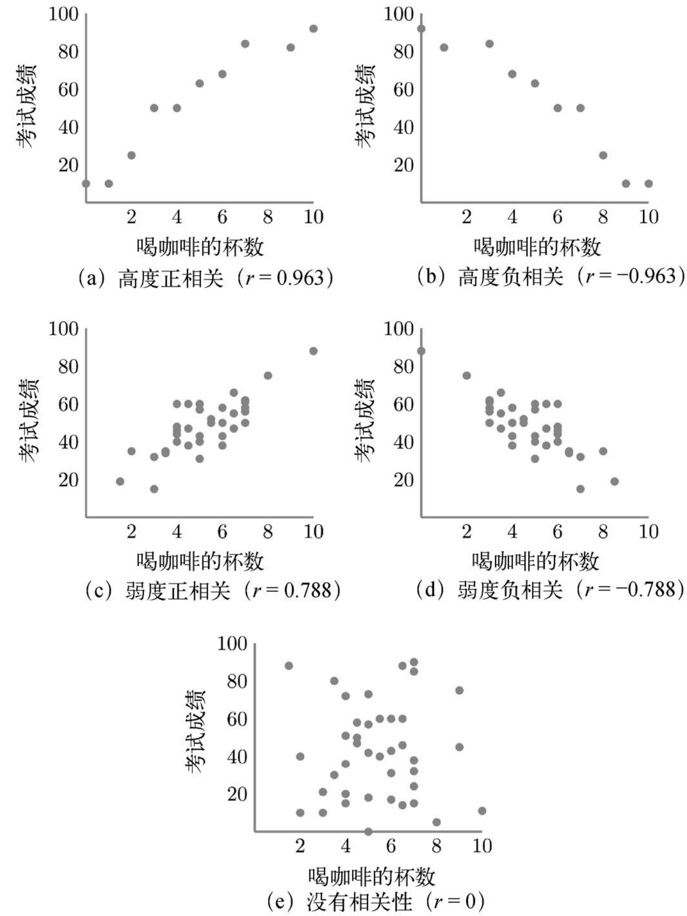


图 3-3　喝咖啡的杯数和考试成绩之间不同强度的相关性

如果将每两个变量之间的关系都变得更弱一些（每次喝同样杯数的咖啡，但考试成绩的变化更大），那么这些黑点就会更为分散，变量之间的相关性也会更低。如图 3-3c 所示，图中的黑点绝大部分仍然呈直线排列，但是偏离中心的距离却要远得多。我们再一次将两个变量之间的关系调转过来（让喝咖啡与更差的考试成绩相关），然后就得到了图 3-3d，两

张图唯一的区别在于一个是上坡面，另一个是下坡面。

注意，当一个变量与另一个变量之间的关系变弱时，要根据喝咖啡的数值来找到考试成绩就难多了，反之亦然。这一点从图上也可以明显地看出来，在前两个例子中，选择一个变量的数值极大地限制了另一个变量可能的数值。然而，如果我们在相关性较弱的情况下，试图预测一个人喝了四杯咖啡后可能会考出的成绩，那么我们的预测将远远不及前面的例子那么准确，因为这时喝四杯咖啡的人考试成绩变化的范围比之前要大得多。变量之间这种不断增加的变化的极限就是变成两个完全不相关的变量（相关系数为零，如图 3-3e 所示），在这种情况下，我们将无法根据饮用的咖啡数量来对考试成绩做出任何预测。

如果我们想知道人们居住的位置和是否开车之间有多强的相关性，应该怎么做？到目前为止，我们介绍的测量相关性的方法一般都用于测量连续值数据（比如股票价格），而不用于测量离散值（比如位置类型或电影类型）。如果我们只有两个变量，而且每个变量只有两个值，那我们就可以用皮尔逊相关系数的简化版——Phi 相关系数。

比如说，我们可以测试人们的居住位置和是否开车之间的相关系数。位置信息要么是市区，要么是郊区或乡下，而开车情况则要么是开车，要么是不开车。和之前一样，我们要测试这些因素是如何共同发生改变的，但此处的“改变”指的是我们看到这两个变量共同出现的频率（而不是这两个数值如何增减）。表 3-1 展示的是数据可能会呈现出的样子。在这个表格中，数据的 Phi 相关系数是 0.81。而我们主要观察的是，测量出来的绝大部分数据是否落在了表格的对角线上。所以，如果绝大部分数值都聚集在“开车/非市区”和“不开车/市区”周围，那么这两个变量之间就存在正相关性。如果绝大部分数值都聚集在另一条对角线上，那么相关性不变，但是相关系数前的符号相反。

表 3-1　居住位置和开车情况的各种组合

	郊区/乡下	市区
开车	92	6
不开车	11	73

然而，相关性强并不一定意味着相关系数也高。皮尔逊相关系数假定两个变量之间是线性关系，即一个变量（比如身高）增大，另一个变量（比如年龄）也会以相同的比率增大。然而，情况并非总是如此，因为还可能存在更为复杂的、非线性的关系。如果不喝咖啡会让人精神不振（并且会降低考试成绩），但是咖啡喝得太多又会让人神经过敏（并且影响考试发挥），那么把我们收集到的一些数据画出来可能就是图 3-4 中的那条曲线。在这个图中，人们喝咖啡的杯数从 0 增加到 5 时，考试成绩是持续上升的，然后在 5 到 10 杯之间，考试成绩随着喝咖啡杯数的增加而慢慢下降。尽管这个案例中的皮尔逊相关系数刚好为零，但是这些数据却呈现出了明显的规律性。很多因果推理方法都很难推理出这种关系，我们将在后面的章节中继续讨论这个问题。鉴于生物医学（比如缺乏维生素或维生素服用剂量过多都可能导致健康问题）和金融（比如将税率和收入联系在一起的拉弗曲线）等应用领域都存在这一问题，所以很值得我们去认真思考一下。

类似地，如果孩子们的体重总是随着年龄的增长而增长，但是体重是以指数级增长的（随着年龄的增长，体重增长得越来越多），那么皮尔逊相关系数会比想象的要低，因为这个指标适用的是线性关系。这就好比我们将数据输入黑匣子，然后不管黑匣子反馈给我们的是什么数字都不假思索地接受，这样是很危险的。在这些相关性被低估甚至看起来是零的案例中，如果我们不进一步研究就直接接受这样的数值，很有可能会错失一些十分有意义的关联。

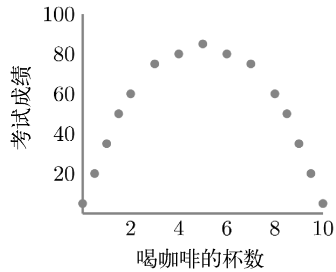


图 3-4　非线性关系（$r = 0.000$）

这就是我们不能把相关系数（不管是皮尔逊相关系数还是其他相关系数）为零理解为不存在任何相关性的部分原因（还有很多其他原因，比如测算中的失误或者导致结果出现偏差的异常值）。另外一个主要原因是，我们所用的数据可能不具代表性，不能反映数据的基本分布情况。如果只使用医院的入院数据和急诊科数据来研究流感致死情况，那我们得到的流感死亡率就会比社会整体人群的实际流感死亡率高得多。这是因为病人一般是因为症状比别人更严重或者还有其他疾病才会去医院（而且去医院的流感病人可能更容易死于流感）。所以我们看到的并不是流感导致的所有结果，而是流感病毒在那些有其他疾病或者流感症状十分严重的病人身上导致的结果。

为了解释限定范围问题，我们假设有两个变量：SAT 总成绩和学习时间。然而，我们并没有所有 SAT 考生的成绩数据，只有那些数学和语文成绩总分超过 1400 分（图 3-5 中的灰色区域）的考生的成绩数据。在这个假设的数据中，成绩好的考生包括那些天生擅长考试的考生（他们不学习也能考得好）和后天刻苦学习的考生。如果仅使用灰色区域的考生的成绩数据，我们是无法找到这两个变量（SAT 总成绩和学习时间）之间的相关性的。但如果我们使用的是所有考生的考试成绩数据，就会发现这两个变量之间存在很强的相关性（灰色区域的考生的学习时间与考试成绩之间

的皮尔逊相关系数为零，而在整个数据集中，二者的皮尔逊相关系数为 0.85）。所以说，我们可以通过以某种结果为限定条件（只研究出现某种结果的案例），然后从毫无关联的变量之间找到相关性。如果 SAT 成绩好且课外活动丰富的学生能够被名校录取，那么仅来自于这些高校的数据则会显示 SAT 成绩和很多课外活动之间存在某种相关性，因为在这个群体中，这两个变量（SAT 成绩好且参加很多课外活动）往往是同时出现的。

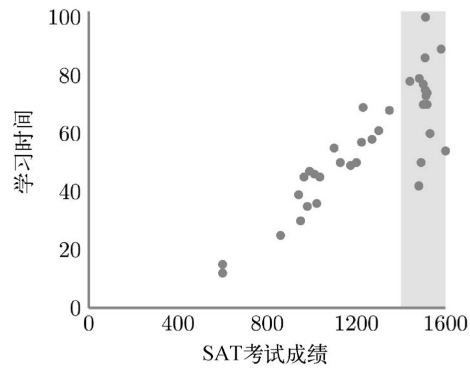


图 3-5　灰色区域的数据代表的是一个限定的范围

这种抽样偏差十分常见，想想那些调查访客政治观点的网站。网站的访客并不是从人群中随机抽取的调查对象。那些带有极端政治偏见的网站的访客，其政治观点与一般人的政治观点之间的偏差就更大了。如果某个网站的所有访客都是现任总统的坚定支持者，那么该网站的调查结果可能会显示，该总统每发表一次重要演说，他的支持率都会上升。但是，这个结论所反映的支持率和重要演说之间的相关性只存在于那些本就喜欢这个总统的人身上（因为接受这个调查的正是这群人）。我们将在第 7 章讨论不同类型的抽样偏差（比如存活者偏差），因为这些偏差会影响我们从实验数据中得出的结论。

有一点需要牢记：我们之所以会找到一些错误的相关性，除了数学方面的原因之外，另一个原因是人们在观察数据时可能会发现一些虚假的规律。有些认知偏差会让我们在无关的因素之间推断出联系，这和抽样偏差相似。比如证实性偏差会使人们去寻找证据来证实他们的观点。如果你认为一种药物会引起某种副作用，那你有可能会去网上搜索其他吃了这个药并且出现了这种副作用的病人。但是，这种做法意味着你是在忽略所有不能证实你的假设的数据，而不是寻找那些有可能让你重新评估你的观点的证据。证实性偏差可能还会导致你对那些与你的假设相矛盾的证据产生怀疑——你可能会认为这些证据的来源不可靠，或者获取这些证据的实验方法有问题。

人们除了在寻找和使用证据时存在偏差，在解释证据时也可能存在偏差。如果一种新药正在接受临床测试，而一名医生已经知道有病人正在服用这种药，并且认为这个药对病人是有帮助的，那么在这种情况下，他就有可能会去寻找迹象来证明这个药物是有效的。由于病人的很多指标都是主观的（比如运动强度和疲劳程度），这就有可能导致医生对这些指标的估算存在偏差，并导致医生推理出一个并不存在的相关性。[13]这个例子来自于一项真实的研究，在这项研究中，发现药物有效的都是那些知情的医生（我们将在第 7 章详细介绍这项研究，并且介绍盲测的重要性）。因此，先验观点不同的人可能会对数据做出不同的解释，从而得出不同的结论。[14]

“错觉相关”是证实性偏差的一种特殊形式，它指的是看到一个实际上并不存在的相关性。关节炎症状和天气之间可能存在着一定的联系，这种联系广为流传以至于人们常常把它当成事实。但是，病人知道这一联系后就有可能会说这两者之间存在相关性，但这不过是因为病人对这种相关性已经有了心理上的预期。然而，当研究人员综合考量了病人自述、临床医生的评价和一些客观的测量数据，试图客观地研究这一相关性时，却发

现这两者之间并没有任何相关性（其他研究人员已经发现，真正的罪魁祸首可能是空气湿度，但是这一结论并不令人信服）。[15] 事实上，当我们把那些关于病人自述的关节疼痛和气压之间关系的数据展示给一些大学生时，他们不仅在没有相关性的时候说看到了相关性，而且在完全一样的序列中找到了正相关性和负相关性。

这种偏差和抽样偏差很相似。我们之所以会错误地认定某种相关性，是因为我们只关注了一部分数据。如果你期望变量之间存在负相关性，那么你就有可能只关注整个数据集中那些能够证实这一观点的一小部分数据。这就是它是一种证实性偏差的原因：人们有可能因为先验的信念而自动将目光投向某些数据。在关节炎与天气的案例中，也许人们对某些证据太过重视（忽视了天气好时关节疼痛的例子，重点突出天气不好时关节疼痛的例子），也许人们看到了一些实际上并不存在的证据（根据他们所预期的联系和天气的变化来讲述不同的症状）。

3.2 相关性的用途

假设我们发现了提交经费申请的时间和是否能够得到资助之间确实存在相关性。提交申请的时间越早，申请书的得分就越高，并且两者之间的相关系数为 1。在这种情况下，如果有人提前一周提交了申请，那我们就可以准确预测出这个人是否能够获得资助，对吗？

很多零售商都在努力寻找能够预测人们购买行为的指标，他们之所以这样做就是依据上面这种逻辑。有人宣称，塔吉特公司在一名青少年的家人还不知情的情况下就已经“知道”她怀孕了。这件事让塔吉特公司一下上了新闻头条。[16] 当然，塔吉特公司并不是真正知道那个女孩怀孕了，而是利用他们从其他顾客身上收集到的海量数据（以及从其他来源购买到的数据）来了解哪些因素与怀孕的各个阶段具有相关性。比如说，经过足

够的观察，塔吉特公司发现单独购买乳液或棉球并不能说明什么，但是那些怀孕的女士通常会同时购买这两样商品以及一些维生素补充剂。在有了足够的采购模式以及预产期（可以从婴儿登记处获得或者根据顾客购买早孕测试纸的日期估算出来）数据后，塔吉特公司就能判断出一名顾客怀孕的可能性有多大，并且能够估算出她已经怀孕多久了。此外，即便我们只知道有人连续购买了两盒早孕试纸，也能从中得知第一张早孕试纸测试结果有可能是阳性的。

Amazon、Netflix 和 LinkedIn 这类网站就是利用相关性来为用户推荐各种互补性商品、用户可能会喜欢的电影和可能会用到的链接。比如 Netflix 网站能够找到那些和你一样喜欢某类电影的人，然后向你推荐一些在这些人中评价很高而你还没有看过的电影。也正是这一点让研究人员在没有用户身份信息的 Netflix 数据集中，能够利用来自另一渠道（比如 IMDB）的数据再次识别用户的身份[17]。我们介绍的只不过是一个基本的构想，真正的算法比这要复杂得多。这些网络公司不一定关心究竟是什么原因让你去做了某件事，毕竟 Netflix 网站能够为你推荐足够多你喜欢的电影，而不用知道你在辛苦工作了一天之后只想看看情景喜剧。

然而，很多基于相关性做的预测都以失败告终，无论这些相关性是否存在对应的因果关系。使用相关性的风险之一在于，对于两个变量之间的任何相关性，我们都可能会找到一些理由来解释这种相关性是如何产生的，从而导致人们对结果过分自信。一个关于数据挖掘的著名案例是，有人利用杂货店的交易数据发现了人们经常同时购买尿布和啤酒这一现象。于是就有人认为，经常在周末来临之前去商店买尿布的男士，会顺便买一些啤酒来“奖励”自己。但是当追踪到这个故事的根源之后，Daniel Power（2002）发现最初的相关性数据并没有提到性别以及时间因素，更没有像有些人说的那样——杂货店特意将这两样商品放得很近，以便一起销售来增加收益。人们在杂货店同时购买的商品也可能只是爆米花和餐巾纸

（晚上要看悲伤的电影），或者鸡蛋和治疗头疼的药物（宿醉）。

假设 Amazon 网站发现，购买某个校园剧和购买 AP 考试（美国大学预修课程考试）复习用书这两个行为高度相关。很明显，美国青少年是这两种购买行为的主体。如果 Amazon 网站只想向同一购买数据群体推荐这些商品,那么他们不知道这两种购买行为的主体也没关系。但如果 Amazon 网站开始把 AP 复习用书推荐给其他国家的顾客，那应该没多少人会买，因为这些考试的参与者主要是美国学生。所以，即便某种相关性既真实存在又十分可靠,如果我们试图将它用在另一个不具备让这一相关性起作用的特征（我们将在第 9 章介绍这样的特征）群体中，那么它可能不会起到任何预测作用。这个相关性并没有告诉我们为什么这些事物之间存在联系——购买者都是十六七岁、正在准备 AP 考试、喜欢看主人公年龄和他们相仿的电视剧——所以我们很难用它来预测其他情况。

这个例子还是相当明确的，但还有一些作用机制很模糊的例子也流传了下来。1978 年，一名体育新闻记者开玩笑似地提出了一个股市新指标：如果美国足联的某个球队赢得了超级碗比赛，那么年底股市就会下跌，否则股市就会上涨。[18] 没有任何理由能使这两件事联系在一起，但是考虑到人们可能用在股市上的各种指标（而且很多时候看起来都是对的），这就足以说服一个没有批判性思维的人去相信这个说法。但是，不了解这一规律的作用机制是什么，就无法预测这个规律什么时候会被打破。

这个规律之所以会起作用，可能是由于人们对这种所谓的相关性的认识影响了人们的行为，因为它已经众所周知了。这也是我们在使用网络搜索或社交媒体上的帖子等观察数据来寻找事物的趋势时，需要关注的一个问题。当用户知道有人在做这些观察时，他们可能会恶意地与系统进行博弈，还可能会改变他们自身的行为（也可能是因为媒体的报道）。

所以，尽管我们可以用相关性来做一些预测，但这些预测有可能会失败，而且我们测量出来的相关性也可能是错误的。

3.3　为什么相关性不是因果关系

有一次我举办了一个关于因果推理的讲座，讲座结束后，一个学生问道："休谟不是说过因果关系实际上就是相关性吗？"这个问题的答案既是肯定的也是否定的。因果关系本身可能更能决定它是不是相关性，但我们却无法确定这一点，只是我们能够观察到的因果关系基本上都是相关性（特殊类型的规律）。然而，这并不意味着因果关系本身就是相关性——只不过相关性正好是我们观察到的关系。这也意味着寻找和分析因果关系的绝大部分工作就是，找一些方法来将具有因果关系的相关性和不具有因果关系的相关性区分开来。

我们可能会通过实验，也可能会通过统计学方法来完成这个工作，但关键是不能在找到相关性之后就停下来。尽管本书讨论了很多关于"表面上的因果关系可能并不是实际上的因果关系"的情形，但是在这一节，我们将简单了解一些在没有对应的因果关系时出现相关性的情形，并在之后的章节中详细阐述其他的情形。

第一个需要注意的问题是：相关性系数是对称的。身高和年龄之间的相关性与年龄和身高之间的相关性完全一样。但是，因果关系可能是不对称的。咖啡让人失眠并不意味着失眠一定会让人喝咖啡（不过这种情况也有可能发生：当人们睡眠不足时，可能会在早上喝很多咖啡）。同样，将反映原因显著性的任何数值（比如条件概率）正着算和反着算也是不一样的。当发现一个相关性时，如果我们完全不知道组成这一相关性的因素的发生顺序，那么每一个因素都可能是导致另一个因素出现的原因（也有可能存在一个反馈循环），而单凭测量相关性并不能区分出这两种（或三种）可能性。如果我们试图用因果故事去解释一对相关因素，就会利用我们的背景知识来推测哪一个因素引起另一个因素的可能性最大。即便性别与中风的概率之间存在相关性，也不可能出现中风决定性别的事。如果我

们发现体重增加和久坐不动的行为之间存在相关性，这两个因素之间的相关程度也并不能告诉我们这一关系的指向性可能会是什么（这两个因素哪一个可能会是原因，哪一个可能会是结果）。

弄错相关性的原因有很多。在 XMRV 病毒和慢性疲劳综合征那个案例中，弄错相关性的原因是实验中使用的样本被污染了。在其他案例中，有可能是计算机程序中的病毒导致的，也可能是誊写结论时的失误导致的，还可能是错误的数据分析方法导致的。表面上的联系可能是统计工具导致的，也可能只是一种巧合，就像股市和足球比赛那个例子一样。然而，还有可能是偏差导致的。既然我们能够从一个有偏差的样本中找到一个并不存在的相关性，那么同样的问题也能导致我们找到一个没有因果关系的相关性。

尽管因果关系能够解释一些相关性问题，但是仍然要牢记这一点：因果关系并不是相关性的唯一解释。比如我们发现按时上班和享用丰盛的早餐之间存在相关性，但是也许这两者都是早起的结果（早起让我们有时间吃早饭，而不是立刻就冲向办公室）。当我们在两个变量之间发现一种相关性时，必须考察一下这种无法测定的因素（一个共同的原因）能否解释变量之间的联系。

在第 4 章的一些案例中，这个共同因素就是时间，我们将会知道为什么我们会在那些随着时间的变化而呈现出一定趋势的因素之间发现很多错误的相关性。如果互联网用户的数量一直在上升，国债的购买数量也一直在上升，那么这两者之间就会出现相关性。但一般情况下，我们指的因素是能够解释相关性的一个变量或者一系列变量。我们可能想知道学习是否能提高我们的成绩，或者那些较好的学生是否更有可能既爱学习成绩又好。也有可能天生的能力是学习时间和成绩的共同原因。如果我们能够改变这种能力，那它可能对成绩和学习时间都会产生影响，然而，任何关于成绩和学习的实验研究都不会对另外两个因素产生影响。

与时间因素相似，相关变量之间不存在直接因果关系的另一个原因是中间变量。比如说，住在城里和较低的体重指数之间存在相关性，因为城市居民走路比开车多，所以活动频率更高一些。所以，住在城里就间接导致了较低的体重指数。但如果搬到城里居住却又开车出行，那这就是一个无效的减肥策略。大部分情况下我们找到的都是间接原因（比如我们找到的是吸烟引起肺癌这一结果，而不是具体的生物进程），但是了解原因具体起作用的机制（原因如何导致结果）能够让我们找到更好的干预措施。

综合数据显示的结果可能会很奇怪。2012 年《新英格兰医学杂志》发表的一篇文章说，人均消费巧克力的数量和每千万人中有人获得诺贝尔奖的人数之间存在显著的相关性，[19] 并且相关系数高达 0.791。在排除掉瑞典之后，这个相关系数提高到了 0.862。之所以将瑞典排除在外，是因为这个国家的数据是异常值，它所产生的诺贝尔奖获得者的数量比人均消费巧克力的数量多得多。要特别注意的是，消费巧克力的数据和获奖数据的来源是不同的，这些数据源分别以每个国家为整体，然后分别进行数据的测算。这就意味着我们并不知道吃巧克力的人和赢得诺贝尔奖的人是不是同一群人。而且获奖人数只是人口总数中极小的一部分，所以获奖人数只要增加几个就能导致相关系数值发生巨大的变化。大多数研究报告都将关注点放在了“吃巧克力和获得诺贝尔奖之间可能存在的因果关系”上：以“巧克力会让人更加聪明吗？”[20]“获得诺贝尔奖的秘诀？多吃巧克力。”[21] 这种标题命名的报告比比皆是。但这项研究并不能证明这些标题中的观点，那些有很多诺贝尔奖得主的国家也可以用很多巧克力来庆祝（记住，相关系数是对称的）。此外，我们无法断言多吃巧克力是否可以增加获奖的机会，或者各国是否应该鼓励公民多吃巧克力，又或者吃巧克力是否是某个因素的指标，比如该国的经济状况。如果还需要进一步的理由才能让你怀疑这一相关性的准确性，不妨想想那些研究人员。他们在没有进行深入分析的情况下就将相关性看成因果关系，甚至通过统计数据发现各个国

家在鹳（一种鸟类）的数量和人口出生率之间也存在显著的相关性，这充分说明了他们的行为有多么愚蠢。[22]

尽管关于巧克力的这项研究有点滑稽，但是这种类型的综合数据经常被用来在某个群体中建立某种相关性，而且由于上述原因，这种数据不易使用又很难解释。将数据与时间联系在一起可能会有所帮助（比如在颁奖之前巧克力的消费数量上升了吗），也有可能是多起事件共同导致了这一变化（比如巧克力的消费数量突然增加，教育政策也发生了变化），并且获奖者通常是在取得能让其获奖的成就之后很久才会获得诺贝尔奖。可能还有很多其他因素也呈现出了相似的相关性，但就在这个巧克力案例之后，又有一个很滑稽的“追踪研究”暗示诺贝尔奖和牛奶之间存在相关性。[23]

3.4 多重测试与 P 值

我们让一位参与者进入功能性磁共振成像扫描仪，然后给这位参与者看各种社会场景的图片，并让其判断每一张图片中人的情感状态。通过功能性磁共振成像扫描仪，研究人员能够测量参与者大脑中各个区域的血液流量，并且经常会用这一测量结果作为神经活动指标[24]，以此来判断不同种类的任务会用到大脑中的哪些区域。最后扫描出来的彩色图像可以向我们展示大脑中哪些区域的血液流量明显增加了，这就是一些研究论文中谈到的大脑中有些区域在特定的刺激下“亮了起来”的含义。找到大脑中那些被激活的区域，可以让我们深入了解大脑的各个部分是如何连接在一起的。

在这项研究中，我们发现参与者大脑中好几个区域的血液流量都发生了十分显著的变化（从统计学上来讲）。事实上，0.05 常常被用作 P 值测量中的临界值（P 值越小证明显著性越高），而大脑中某个区域的活跃度的 P 值只有 0.01。[25]那么，这个区域是否和人们想象他人情感（换位思考）的活动有关呢？

如果这项研究的参与者是一条死掉的三文鱼，上述结果发生的可能性似乎不大。一条死鱼怎么能对视觉刺激做出反应呢？上述结论无论使用什么样的常规临界值，报告中的显著性都会非常高。所以，问题不在于这一显著性是否被夸大了，而在于它是如何出现的。为此，我们需要简单地插入一些统计学知识。

研究人员常常需要确定某种效果是否具有显著性（某种相关性是真实存在的还是统计假象），或者两个群体之间是否有差异（人们在看人和看动物时，大脑的活跃区域相同吗），但他们需要一些定量的指标来客观地确定哪些发现是有意义的。P 值就是一个用来测量显著性的常用指标，人们用它来对比两个不同的假设（零假设和对立假设）。

P 值告诉我们，如果零假设成立，那我们看到一个至少和已经观察到的结果一样极端的例子的概率有多大。

对于我们而言，这些假设可能是指两个事物之间没有因果关系（零假设）或者有因果关系（对立假设）。或者是另一种情况，零假设可能认为硬币是均匀的，而对立假设则认为硬币是不均匀的。人们常常把 P 值误解为零假设成真的概率。尽管人们一般把 0.05 作为临界值，但是并没有任何定律规定 P 值在 0.05 以下的结果就一定是显著的，而 0.05 以上的就一定是不显著的。这只是一种惯例，而且选择 0.05 作为临界值也极少会遭到其他研究人员的反对。[26]这些数值并不一定能完全反映实际的显著性，因为显著性极小的结果可能会有极小的 P 值，而显著性极大的结果都可能达不到统计学对显著性 P 值的要求。

在电影《罗森・格兰兹与吉尔・登斯顿之死》的开头，有几个人在掷一枚刚刚捡到的硬币，他们越掷越奇怪，因为这个硬币每次都是正面朝上，一连 157 次都是如此。[27]一枚硬币连续 157 次正面朝上的概率极小（准确地说只有 $1/2^{157}$），而与之同样极端的情况就是连掷 157 次都反面朝上。所以，罗森・格兰兹与吉尔・登斯顿观察到的确实是一个 P 值极低的

事件。但这并不意味着一定有什么奇怪的东西在作祟，只不过这一结果在硬币均匀的情况下不大可能出现而已。

再看一个没那么极端的情况，比如我们连抛了 10 次硬币，其中 9 次是正面朝上，1 次是反面朝上。这个结果（零假设为硬币是均匀的，对立假设为硬币正面朝上或反面朝上都是不均匀的）的 P 值是 9 次正面朝上和 1 次反面朝上的概率，加上 9 次反面朝上和 1 次正面朝上的概率，加上 10 次正面朝上的概率，加上 10 次反面朝上的概率。[28]之所以要加上全部正面朝上的概率和全部反面朝上的概率，是因为我们计算的是至少与我们观察的事件同样极端的事件的概率，而全部正面朝上和全部反面朝上这两个事件比 9 次正面朝上和 1 次反面朝上这样的事件更为极端。这个案例的对立假设为这个硬币是不均匀的，既不是仅仅偏向正面，也不是仅仅偏向反面，这就是我们要把连续出现反面朝上的案例也包括进去的原因。图 3-6 的柱形图展示的是 10 枚硬币每个掷 10 次，其中正面朝上的次数。如果每一枚硬币抛出的结果都恰好是 5 个正面和 5 个反面，那么在这些横轴为 10 的柱形图中，每个图形的竖形柱都会集中在 5 这个中心点上。但在实际生活中，硬币抛出的结果既会出现数值大于 5 的情况，也会出现数值小于 5 的情况，甚至还出现了一种全部反面朝上的情况（由图中最左边的小竖形柱来表示）。

即使我们用的是一枚均匀的硬币，上面这个事件出现的概率依然很小，但如果我们抛 100 枚均匀的硬币又会出现什么结果？实验的次数多了，我们就有更多的机会碰巧看到一些似乎很反常的事情。比如说，每个人买彩票中奖的概率都非常低，但如果买彩票的人足够多，那我们几乎可以保证总会有人中奖。图 3-7 所示的柱形图和前面的柱形图一样，但这一次用的硬币是 100 枚而不是 10 枚。在这种情况下，如果没有看到任何一枚硬币抛出了 9 个或 10 个正面或反面，那我们会感到更加惊讶（同样，如果彩票的中奖率为千万分之一，但是有 1 亿人买了彩票却没有一个人中奖，这也会让人感到惊讶）。

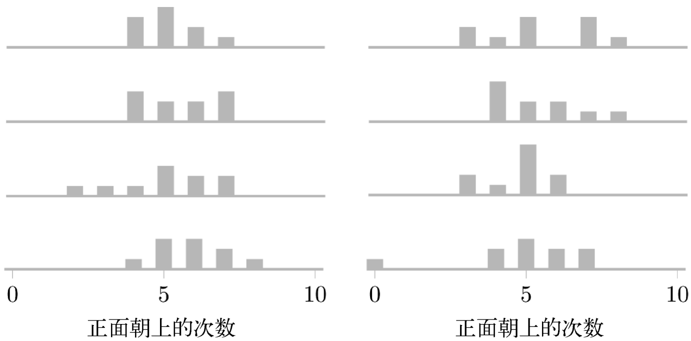


图 3-6 一个柱形图表示一次实验的结果，每次实验都是 10 枚硬币各枚掷 10 次。根据硬币正面朝上的次数，每掷 10 次就能获得图形上一个数据点。图中展示的是 8 次实验的结果

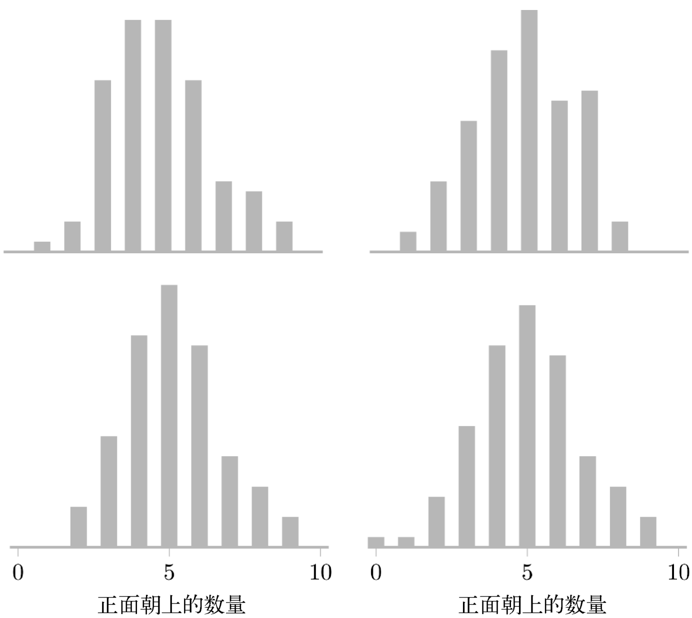


图 3-7 100 枚硬币每枚掷 10 次的结果。图中展示的是 4 次实验的结果

一次进行多个测试是会出现问题的，这一问题正是我们一开始介绍的功能性磁共振成像研究中出现的问题。在磁共振成像研究中，人们考察了大脑中好几千个很小的区域（在研究人脑时所考察的小区域数量更多，因为人的大脑要大得多），所以其中有一个区域呈现出明显的血液流动现象也并不奇怪。这种问题被称为多重假设检验，顾名思义，它指的是同时检验大量假设。随着能够产生海量数据集的新方法（比如功能性磁共振成像技术和基因表达阵列）以及“大数据”的出现，多重假设检验的难度越来越大。以前我们可能只能用一个实验检验一种假设，但现在我们能够分析上千个变量。由于检验的数量庞大，所以即使发现有些变量之间存在一些相关性也并不应该感到奇怪。

在那个以三文鱼为参与者的研究中，研究人员检验了数千个假设，每个假设都认为大脑中的某个区域会在实验任务中表现出显著的活跃性。这项研究的目的实际上就是告诉人们，这些测试可能会单纯因为巧合而出现一些似乎具有显著相关性的结果。这一研究还介绍了一些纠正多重对比问题的统计方法（基本上每一次测试都需要使用更为严格的临界值）。使用了这些统计方法后，即便我们放松对 P 值的要求，也不会再出现显著的活跃性。[29]

要牢记一点：在阅读关于显著性发现的报告时，如果这个发现是从大量同时进行的测试中计算出来的，那么就有必要看看这些报告的作者们是如何处理多重对比问题的。至于究竟该如何（以及何时）纠正这个问题，统计学家们意见不一，但从总体上来说，这些分歧归根结底是要确定哪一种错误的影响更大。在纠正多重对比的过程中，我们其实是选择了减少错误的发现，即使因此错过了一些重要的发现（导致漏报）也没关系。而如果我们认为不应该纠正多重对比问题，则是选择了宁愿找到一些错误的发现，也不愿错过一些真正正确的发现。

这两种错误一直是一个此消彼长的问题，究竟哪一个更合适则完全

取决于人们各自的目的。[30]对于那些探索性分析来说，分析结果会继续接受实验的验证。在这种情况下，我们可能想让分析结果包括的范围更广一些。相反，如果我们正试图为一个昂贵的药品开发项目挑选一个针对性很强的参与者群体，那么每一个错误的推理都可能导致我们浪费大量的金钱和努力。

3.5　没有相关性的因果关系

尽管我们常常讨论为什么某个相关性不是因果关系这一问题，但我们也必须承认，有些因果关系中确实没有明显的相关性。这意味着仅靠相关性并不足以证明因果关系的存在，而且相关性也并不是因果关系的必要条件。辛普森悖论就是个例子（我们将在第 5 章详细讨论这个案例）。即使两个事物在一些小群体中存在某种联系（比如说与某个人群当前使用的治疗方案相比，某种试验药能够改善治疗效果），但当我们将这些小群体合在一起时，可能就会发现二者之间不存在任何关系，或者存在完全相反的关系。如果某种新药的使用者往往是那些病得很重的病人，而病得不重的病人往往会使用当前的治疗方法，那么在不考虑病情严重程度的情况下，试验药在整个病人群体中导致的结果似乎要更严重一些。

再举一个没有相关性的因果关系：长跑对体重的影响。虽然长跑能够消耗热量从而减轻体重，但是长跑也能导致食欲大增从而增加体重（而这又会对减肥造成负面影响）。根据每种影响的强度不同，或者根据调查的数据不同，跑步的积极作用可能恰好会被它的消极作用抵消，结果人们就会发现在跑步和减肥之间不存在任何相关性。这个例子的因果结构如图 3-8 所示。还有一个关于吸烟的例子：有些吸烟者会加强锻炼并改善饮食，以此来抵消吸烟对他们健康的负面影响，最后导致人们无法找到吸烟在某些方面对他们的影响。在这两个案例中，同一个原因通过不同的路

径既对人们产生了积极的影响，也带来了消极的影响。这就是为什么我们可能观察不到任何相关性，或者只能观察到极弱的相关性（记住，测量本身并不是完美的）。

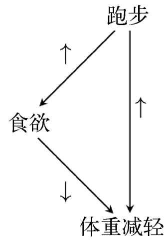


图 3-8　积极的因果关系（向上的箭头）与消极的因果关系（向下的箭头）。根据人群的情况，这些影响可能会相互抵消

我们已经研究了一些可能会导致我们无法发现某个相关性的其他原因（比如抽样偏差、变化量不足、证实性偏差、非线性相关，等等），而且也经常听说相关性并不意味着因果关系。但这句话倒过来也很重要：因果关系并不总意味着相关性。[31]

注释

1. Lombardi 等（2009）。
2. 这样的研究和理论有很多，Afari 和 Buchwald（2003）写了一篇评论文章，讨论了其中的一些研究成果和理论。
3. Holgate 等人（2011）的研究简短地介绍了人们在研究 CFS 的过程中遇到的各种困难，包括各种定义上的差异。
4. 有些研究未能成功复制 CFS/XMRV 联系，其中包括 Erlwein 等（2010）和 van Kuppeveld 等（2010）的研究。
5. Lo 等（2010）。
6. 第二篇即将发表的文章被作者撤回了（Lo 等，2012），而 Mikovits 研究团队的文章首先被团队部分成员撤回了一部分内容（Silverman 等，2011），后来又被《科学》杂志完全撤回了（Alberts，2011）。

7. 其他研究团队向人们解释了为什么这些结果可能是由于 XMRV 污染造成的，并且通过将另外两种病毒结合在一起来推断出这种病毒实际上来源于实验室。人们在《逆转录病毒》上发表了四篇专门研究这一污染问题的文章（Hué 等，2010；Oakes 等，2010；Robinson 等，2010；Sato 等，2010），后来又有人发表了一篇讨论 XMRV 的来源的文章（Paprotka 等，2011）。
8. Cohen（2011）。
9. Alter 等（2012）。
10. 皮尔逊相关系数（由 Karl Pearson 提出）从数学角度被定义为：

$$r = \frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma(X - \bar{X})^2 \Sigma(Y - \bar{Y})^2}}$$

. 在这个公式中，X表示平均数。注意，在分子中，我们将某一测量点 X 与平均值的差和 Y 与平均值的差的乘积累加在一起。在分母中，我们计算的是个体的变化量。
11. 在皮尔逊相关系数中，我们需要除以变量标准差的乘积。因此，如果这两个标准差中有一个为零，这个系数由于要除以零就无效了。
12. 比如说，Salganik 等（2006）曾指出，让那些歌曲继续流行下去的方式是不可预测的，他们还证实了流行歌曲的成功并不完全取决于歌曲的质量。想要了解更多这方面的信息，参见 Watts（2011）。
13. Noseworthy 等（1994）。
14. 想要阅读更多关于其他认知偏差的信息，参见 Tversky 和 Kahneman（1974）。
15. Patberg 和 Rasker（2004）；Redelmeier 和 Tversky（1996）。
16. DuHigg（2012）。
17. Narayanan 和 Shmatikov（2008）。
18. Koppett（1978）。
19. Messerli（2012）。
20. Pritchard（2012）。
21. Waxman（2012）。
22. Höfer 等（2004）；Matthews（2000）。
23. Linthwaite 和 Fuller（2013）。
24. Heeger 和 Ress（2002）。
25. Bennett 等（2011）。
26. Fisher（1925）一开始曾暗示 0.05 可能是一个很有效的临界值，但他并不建议所有人在任何情况下都应该使用 0.05 作为临界值。

27. Stoppard（1990）。有趣的是，硬币连续出现正面朝上的次数与最初投掷的结果相比增加了。
28. 这里的 P 值是 0.022。因为出现 10 次正面朝上（或反面朝上）的概率是 0.001，而出现 9 次正面朝上（或反面朝上）的概率是 0.01，将这些数值加在一起正好是 0.022。
29. 想要详细（且专业地）了解如何调整以便进行多重假设检验，参见 Efron（2010）。
30. 有观点认为，我们不应该调整，而应该进行多重对比。想要了解这个观点的更多信息，参见 Rothman（1990）。
31. 我们将在第 6 章更加深入地讨论这个问题，并且探讨这些所谓的违背忠实原则的行为是如何影响我们通过计算推理原因的能力的。

第 4 章　时间

时间如何影响我们感知因果关系和进行因果关系推理的能力？

2001 年，研究人员做了一组随机对照实验来检测祈祷是否会提升病人的治疗成效，比如缩短病人住院的时间。[1] 这个双盲实验（医生和病人都不知道谁在哪一组）召集了 3393 名血流感染的成年住院病人，其中大约一半分到了控制组，还有一半分到了祈祷干预组。测试结果显示，干预组病人的住院时间缩短了，发烧程度减轻了，而且与控制组相比，这两个因素的变化程度在统计学上都具有显著性（P 值分别为 0.01 和 0.04）。

然而，既然这种干预措施如此有效，那它为什么没有被所有医院采用呢？其中一个原因是，这项研究中的病人的住院时间是在 1990—1996 年，这意味着康复祈祷是在他们住院且治疗结果出来之后很久才发生的。实际上，祈祷不仅是在结果出来之后才发生的，而且祈祷发生的地点和时间距离病人住院的地点和时间十分遥远，为病人祈祷的人也从未和病人有过任何接触。

当下的一个原因影响了过去发生的事情，这与我们对因果关系的认识完全相反。一般情况下，原因的出现往往要先于结果（即便原因和结果在时间上不是很接近），而且原因和结果之间会存在一定的物理联系。这项研究是按照随机实验的常规标准（比如双盲）进行的，而且实验结果在

统计学上具有显著性。这篇文章吸引了很多读者，他们纷纷来信，谈到了这篇文章的哲学与宗教意义。但是，问题的重点不是信仰。相反，这项研究向读者提出了挑战：如果这些结论来自一项符合他们标准的研究，而且研究方法合理，研究结果在统计学上也具有显著性，那么他们会接受与其先验信念极为矛盾的结论吗？

“当下的某个原因能够导致某件事情在过去发生”，你能想象到有哪个研究能够说服你相信这种理论吗？这项研究看似合理，但是我们却不大可能相信这一结果是干预措施导致的，因为它违反了我们对因果关系中时间因素的认识。如果你对一个假设的先验信念足够接近实践活动，那么任何实验可能都不会真正改变你的想法。

尽管事件发生的顺序对因果关系至关重要，但是我们也十分在意原因和结果之间的延迟。如果你和一个得了流感的朋友一起看电影，三个月后你也得了流感，那你可能不会认为是你朋友传染给你的。但如果你认为接触流感病人就会染上流感，那你为什么不把责任推到你朋友身上呢？因为并不是接触了某个病毒就会生病，而是由于病毒存在潜伏期，接触某个病毒并不会立即引发相应的症状，而且也不会导致人们在很久以后才出现流感症状。实际上，接触病毒和引发疾病之间的时间很短，我们可以利用这个时间段来缩小范围，找到可能是哪一次接触引发了某个疾病。

时间因素往往能让我们区分原因和结果（体重下降之前就生病了，这说明这个疾病不可能是体重下降引起的）、能让我们的干预措施发挥作用（有些药物必须在接触病毒后立刻服用），还能让我们预测未来将会发生的事件（知道股票价格的上涨时间比仅仅知道它会在未来某个不确定的时间段上涨更有用）。但是，时间可能也会造成误导，因为我们可能会在毫无关联的、具有相似趋势的时间序列中找到相关性。当结果出现延迟时

（比如接触的环境与健康状况），我们可能会找不到导致这一结果的原因。当一个事件经常发生在另一事件之前时（卖伞的小贩们会在下雨之前开始卖伞，但这绝对不是下雨的原因），我们可能会错误地将一些无关的事件联系在一起。

4.1　因果关系的感知

我们是如何从“运动与减肥之间的相关性”推理出“是运动导致体重下降而不是体重下降导致运动”这样一个结论的呢？相关性是一种对称关系（身高和年龄之间的相关性与年龄和身高之间的相关性完全相同），但因果关系却是不对称的（炎热的天气会让人跑步的速度放慢，而跑步却不能引起天气的变化）。我们可以根据背景知识了解到人们跑步的速度是不可能影响天气的，但在从相关性到提出因果假设这一过程中，最关键的信息之一就是时间。

休谟处理非对称性问题的方法是，默认原因和结果不可能同时发生，而且原因必须先于结果发生。因此，如果我们观察一些正常发生的事件，那一定是先发生的事件导致后发生的事件。[2] 然而，休谟的哲学研究主要是理论性的，虽然从直觉上来讲，我们依靠时间上的优先性来感知因果关系是没问题的，但这并不意味着事情就一定是这样的。

如果我们看到一个台球向另一个台球滚动并且撞击了它，然后第二个台球开始向前滚动，我们会自然而然地认为第二个台球的运动是第一个台球引起的。如果第二个台球被撞击后过了很长时间才开始滚动，或者第一个台球并没有直接撞上第二个台球，而是在离第二个台球不远处就停住了，那么你可能就不大会认为第二个台球的运动是第一个台球引起的了。是事件发生的时间导致了人们对因果关系的感知吗？还是这种感知取决于空间上的位置？

为了弄明白这一点，我们再次回到第 2 章提到的心理学家 Albert Michotte 的研究中。20 世纪 40 年代，Michotte 做了一系列实验来弄清楚时间和空间是如何影响人们对因果关系的感知的。[3] 在一个经典的实验中，参与者看到两个影子在屏幕上移动，然后他们要描述自己都看到了什么。Michotte 试图通过改变影子的运动特征，比如这两个影子之间是否有接触，一个影子的运动是否先于另一个影子，从而确定是哪些特征导致参与者产生了两者之间具有因果关系的印象。

在因果关系感知研究中，Michotte 的研究影响深远。当然，他的研究在研究方法和研究结果的证明上也存在一些争议。很多时候，我们并不清楚某项研究的参与者到底有多少、他们的人口特征是什么、他们的反应从何而来、他们是如何被挑选出来的，也不知道这些参与者的具体反应到底是什么，以及为什么他们的这些反应会被看成是有因果关系的。据 Michotte 称，这些参与者很多都是同事、合作者和学生——由这些参与者组成的群体比整个人口群体的专业性要高。尽管 Michotte 的研究为将来的实验研究提供了一个重要的起点，但他的研究结论还需进一步地复制和追踪。[4]

在 Michotte 的实验中，有两个影子从屏幕上经过，这两个影子没有发生任何接触且同时开始运动（如图 4-1a 所示），这时参与者往往不会用因果关系来描述影子的运动。[5] 在另一个实验中，一个影子朝另一个影子运动，然后第二个影子在接触了第一个影子后也开始运动（如图 4-1b 所示）。在这种情况下，参与者通常会认为是第一个影子引起了第二个影子的运动，[6] 并会使用一些表示因果的语言（比如推动、发动等）来描述两者的关系。这些场景只是描绘了影子在屏幕上的运动过程，它们的运动轨迹之间并没有真正的因果依赖性，但人们依然会用因果关系来解释和描述整个运动过程。[7] 观察者认为第二个影子的运动是由第一个影子发起的，并将第一个影子看成一个发动器，这种现象被称为发动效应。值得注意的是，在两个影子之间加入空间距离（如图 4-1c 所示）并不能消除人们认为

它们之间存在因果关系的印象。[8]如果事件发生的顺序不变，一个影子朝着另一个影子运动，在碰到另一个影子之前停了下来，然后另一个影子在第一个影子停下后立刻开始运动，那么参与者仍然会使用一些表示因果的语言来描述这一过程。从这个实验可以看出，有些情况下，与空间邻近性相比，时间上的优先性可能是一个更为重要的信号。当然，这也要看问题本身的特点以及事物之间的空间距离到底有多大。

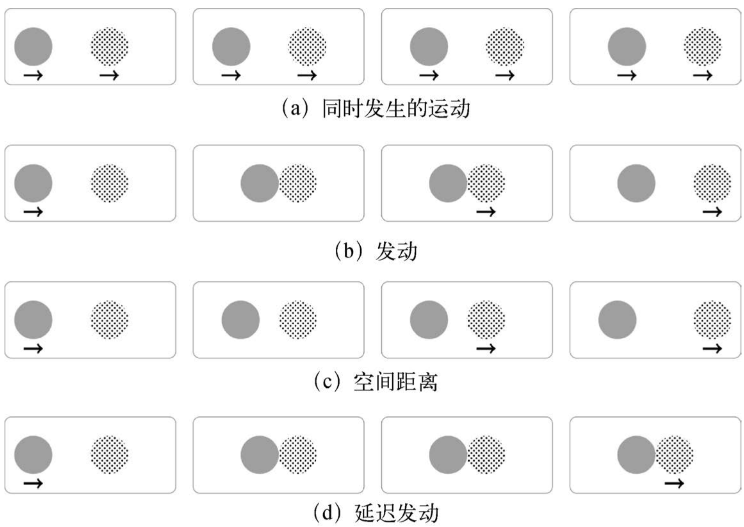


图 4-1　上面几张图展示的是 Michotte 所做的各种实验中的几种。在这几种实验中，影子以不同的方式运动。箭头表示这个影子正在运动以及它运动的方向

尽管我们并不能根据发表出来的描述文字准确地复制这个实验当初所用的方法，但还是通过其他研究活动证实了发动效应。但我们证实的发动效应的普遍性比 Michotte 暗示的低，可能只有 64%~87%的观察者在第一次看到一个运动时会用含有因果关系的语言来描述这个运动。[9]

假设一个球正在滚向另一个球。第一个球一接触到第二个球就停了

下来，在短暂停顿之后，第二个球按照第一个球的运动方向开始滚动。第二个球的运动是第一个球引起的吗？停顿的时间是 1 秒还是 10 秒很重要吗？休谟认为时空上的邻近性对因果关系的推理十分重要，但在实践中，我们无法看到每个因果关系链中的每一个环节。为了考察延迟是如何影响人们对因果关系的判断的，Michotte 设计了一些和上述实验一样的场景，在一个影子结束运动之后，另一个影子过一段时间再开始运动，如图 4-1d 所示。他发现尽管两个影子之间的距离十分接近（这些影子确实有接触），但时间上的延迟消除了人们认为这两个运动之间存在因果关系的印象。[10]

除了参与者的专业水平（以及参与者对实验和 Michotte 提出的假设的了解程度）外，这些实验还有一个局限性，就是参与者只是描述了这些影子在屏幕上的行为，而没有试图通过与这种行为进行互动来发现系统的各个特性。描述与互动的差异就好像一个是看别人按电梯按钮然后等电梯什么时候会来，一个是按照自己的时间来选择什么时候按电梯按钮。Michotte 的研究告诉我们，人们在特定情况下可能会用表示因果关系的语言来描述一些场景，但在一个物理系统中，如果参与者能够控制原因发生的时间，又会出现什么情况呢？

在 Michotte 的研究基础上，Shanks、Pearson 和 Dickinson 也做了很大的贡献，他们研究了时间对因果关系判断力的调节作用，而且将系统变成了参与者之间互动的工具。在这个实验中，按下键盘上的空格键，电脑屏幕上就会出现一个闪烁的三角图形，而参与者必须判断按空格键的行为在多大程度上导致了这个三角形的出现。

研究人员将按空格键和出现三角形这两件事情之间的延迟时间从 0 秒延长到了 2 秒，然后发现这种延长导致参与者认为空格键引起三角形出现的可能性变小了。研究人员又使用一系列的延迟时间（从 0 秒到 16 秒）进行了试验，然后他们发现，平均来说，按空格键的行为与出现三角形的效果之间的延迟越长，参与者认为这两者之间存在因果关系的可能性就越低。

在使用实体对象做实验时，如果在两个物体发生接触后，其中一个物体隔了很长时间才开始运动，那么我们完全有理由怀疑这个物体的运动并不是另一个物体引起的。但在其他情况下，我们不应该指望某个结果会立即出现。接触一个病原体并不会立即让人生病，干预政策可能需要很多年才能产生可衡量的效果，通过运动来减肥是一个缓慢的过程。这些实验似乎表明，原因和结果之间的延迟会减少人们对因果关系的判定，或者导致人们做出错误的推理，这样的结论似乎就存在问题了。

最近有研究发现，尽管原因和结果之间的延迟会增加人们准确判断因果关系的难度，但这在一定程度上也可能取决于人们对这种延迟的不同预期。如果在击打高尔夫球和高尔夫球飞出去之间出现了 10 分钟的延迟，这就与我们所知道的物理知识严重矛盾。但如果一个人接触了致癌物，然后在十年之后才得了癌症，这么长的延迟却不会让人觉得意外。延迟的长度对我们的影响可能有一部分取决于我们对问题的认识，以及我们对事物运行机制的了解。在目前提到的心理实验中，有很多心理实验的设置总能让参与者想起一些熟悉的场景。在一些场景中，他们预计某个原因会立即引发某种结果。比如说，Michotte 移动的圆圈代表一些圆球（在这个实验中，人们认为第二个球在被撞击之后会立即滚动起来，而撞击和滚动之间的任何延迟都是异常的），而 Shanks 等人在研究中用的则是键盘（在这个实验中，人们预计在按下空格键之后，电脑屏幕很快就会出现反应）。如果我们给参与者一些场景，比如让参与者判断吸烟是否是某个病人得癌症的原因，然后告诉参与者某个人的吸烟历史和肺癌的诊断结论，那么参与者有可能会发现，一个一周前开始吸烟的人在一周后被诊断为癌症的可能性极小，因为吸烟可能需要更长的时间才能引起癌症。

为了研究这个问题，Buehner 和 May 做了一个与 Shanks 等人的研究类似的实验。但在这个实验中，Buehner 和 May 给了参与者一些背景知识，

告诉他们在按下按键和屏幕上出现三角形之间可能会有延迟，由此操纵了参与者对延迟的预期。参与者被分为两组，其中只有一组参与者（实验组）提前被告知在按按键和屏幕上出现三角形之间可能会有延迟。在对比了两组实验结果之后，我们发现尽管时间上的延迟总会导致人们降低对因果关系的判定指数，但这种延迟对实验组的影响要小得多。此外，实验的顺序（参与者是先看到有延迟的效果，还是先看到没有延迟的效果）也会对实验结果产生显著的影响。如果参与者先看到的是有延迟的效果，那么他们感受到的因果关系的强度要比先看到没有延迟的效果高得多。由于实验顺序不同而产生的这一影响表明，影响我们判断的不仅是事件发生的顺序或者事件之间的延迟长度，还有这些因素和先验知识的相互作用。在 Michotte 的实验中，参与者看到圆圈在屏幕上移动，但是他们对这些圆圈的解释却好像它们是实体对象一样，因此他们的解释中还包括他们对动力传递的预期。

在 Buehner 和 May 的研究中，参与者提前被告知的信息限制了时间上的延迟对因果关系判断的影响，但即使参与者已经知道会存在这种滞后，它依然会影响他们对因果关系的判断，这就很奇怪了。之所以会出现这一现象，可能是因为实验内容仍然包括按键后屏幕上就会出现某种效果。也有可能是因为人们对计算机处理输入指令的反应速度已经有了强烈的预期，无法通过预先说明来消除这种预期。即使参与者已经提前知道可能会存在延迟，他们还是会利用先前对按键后的屏幕反应速度的预期来进行因果判断。

后来，研究人员用一个节能灯的例子（参与者可能都遇到过从按开关到亮灯之间的延迟）成功消除了延迟对人们判定因果关系的负面影响。在这个实验中，那些被告知可能会有延迟的参与者无论是在有延迟的情况下还是在没有延迟的情况下，他们对因果关系判定的平均值都是一样的。[11]

在上述两种情况下，虽然延迟已经不再影响我们对因果关系的判断，但参与者依然认为即时效应是由某个原因导致的。即便他们所得到的一些与问题有关的信息并不支持这样的结论，他们也依然这样认为。我们所面临的挑战之一是，要设计一个能够保证参与者对延迟的长度有着强烈预期的实验，并且还要保证这些预期与他们先前对事物的作用机制的认识相一致。有一个实验利用了一个倾斜的盘子，让一颗弹珠从高处进入盘内并一直向低处滚动直到见底，然后去触发盘子底部一个控制灯光的开关。盘子的角度可以调整。如果盘子几乎是垂直于地面的，那么在弹珠进入盘内和灯光亮起来这两件事情之间几乎不可能出现延迟；如果盘子几乎是平行于地面的，那么这两件事情之间就很有可能出现延迟。这和第 2 章提到的心理学实验中所使用的快与慢的机械装置相似。通过这种设置，Buehner 和 McGregor 证实了这一点：有些情况下，即时效应可能会降低一个原因的可信度。之前的大部分研究都表明，延迟增加了寻找原因的难度，即便没有增加难度，最多也只是不影响推论活动。但 Buehner 和 McGregor 的研究表明，在有些情况下，延迟居然会对寻找原因的活动有所帮助（延迟短和盘子的倾斜度比较低这两个因素会降低两个事件之间存在因果关系的可能性）。这一发现至关重要，它表明延迟并不总是会妨碍我们的推理活动，也不总是会降低原因的可能性。就时间问题而言，最重要的是我们观察到的延迟与我们预期的延迟之间的关系。

这些实验中需要注意的主要问题是，按下按键后会在多大程度上引发视觉效果，或者是否是弹珠让灯亮了起来，而不是去辨别多种可能的原因。通常情况下，我们不仅需要判断一个特定事件引起某种结果的可能性有多大，而且还要假设哪些因素可能会是首要原因。如果你食物中毒了，那不仅要考察某个食物是不是中毒的原因，还要考察你所吃过的所有食物来判定罪魁祸首。这时候，时间可能会是一个重要的线索，因为上周吃的

食物不太可能是罪魁祸首，而最近吃的食物则更有可能是中毒的原因。

有些心理学研究为这种类型的推理提供了依据。这些研究表明，在因果关系未明的情况下，时间信息可能确实会比其他线索（比如这些事件同时发生的概率有多大）更加重要。然而，这也可能会导致我们推理出错误的结论。在食物中毒的案例中，你可能会仅依据时间因素就将最近吃的东西错误地当成罪魁祸首，而忽视了其他信息，比如哪些食物或者哪些饭店与食物中毒的联系最为密切。Lagnado 和 Sloman 所做的一项研究表明，即便我们告知参与者可能会有延迟，这些延迟可能让他们观察到的各个事件之间的顺序不那么可靠，这些参与者依然会根据一些因果联系得出错误的结论。即便参与者观察到的这些因素共同出现的次数与时间信息矛盾，他们依然会依赖时间信息来发现各种关系。

假设你在按一个开关。你不太清楚这个开关是干什么的，所以你按了它很多次。有时你一按开关就有一盏灯亮了，但有时要过一会儿灯才会亮。有时这中间会有 1 分钟的延迟，有时会有长达 5 分钟的延迟。是这个开关打开了灯吗？这有点像按人行道过街按钮的结果，按下按钮似乎并不会让信号灯快点切换。很难判断二者之间是否存在因果关系，因为按按钮和信号灯切换之间的时间间隔变化太大。关于改变延迟稳定性的实验表明，如果原因和结果之间的滞后情况稳定（比如“三角形总是在按下按键 4 秒之后出现在屏幕上”与“三角形总是在按下按键 2 秒到 6 秒之后出现在屏幕上”），那么人们对因果关系的评分就会比较高，随着时间间隔的变化范围不断扩大，因果关系的评分也会不断变低。[12] 从直觉上来说，如果时间间隔在平均值上下略微浮动，那么很有可能是其他因素的细微变化或者是观察过程中的延迟造成的。相反，如果时间间隔的变化范围巨大，比如一种药物在服药后的 1 天到 10 年出现副作用，那么很有可能还存在其他能够决定时间间隔（加快或延迟某种结果出现的时间）的因素，而且可能有不止一种原果机制或存在某种混乱的关系。

4.2　时间的方向性

假设有个朋友跟你说某种新药对她的过敏有效。如果她说这种新药让她不打喷嚏了，那么你会怎样假定开始吃药和不打喷嚏之间的顺序呢？根据我们暗示的这种关系，你很可能会认为吃药在前，停止打喷嚏在后。事实上，时间可以帮助我们寻找事件发生的原因，它和因果关系之间的紧密联系也能让我们从因果关系中推理出关于时间的信息。有些研究发现，关于原因的了解会影响我们感知事件之间时间间隔的方式，[13]甚至还会影响我们感知事件发生顺序的方式。[14]

有时两个事件看起来好像是同时发生的，但其实是测量粒度或观察能力有限导致的。比如，微阵列实验一次检测数千个基因的活动情况，而对基因活动水平的检测通常是按固定的时间间隔（比如每小时一次）进行的。从数据上看，两个基因的活动模式看起来好像是一样的——同时出现过度表达或者低表达。然而，事实可能是那个被上调的基因引起另一个基因随即也被上调。但是，如果我们看不到这种排序，而且也没有任何背景知识表明肯定有一个基因先发生了变化，那么我们能确定的只是这两个基因的表达水平是相关的，而无法确定一个基因是否会导致另一个基因被调节。

同样，病历所记录的并不是每个病人每天的数据信息，而是一系列不规则的时间点（病人去看医生的时间点）的数据信息。因此，我们可能会发现某个病人在某个时期既在服用某种药物，又在忍受着某个副作用。但是，我们只知道这两件事情都发生了，却不知道病人是否是先服药后出现的副作用，也不知道这个药物是否是引起这个副作用的潜在原因。在一些长期队列研究中，参与者可能每年才接受一次调查，所以如果环境暴露或者其他因素在短期内对参与者产生了某种影响，那么这一因果序列是无法通过这种长期的队列研究而被发现的（假设这些事件可以被准确地回忆起来）。在很多情况下，两个事件中的任何一个事件都有可能先发生，而

它们共同出现并不代表两者之间存在某种特定的因果顺序。

没有任何时间信息的情况是最复杂的，比如在横断面研究中，所有数据都是在同一时间收集的。某个横断面研究调查了某个人群中的任意一个小群体，以此来判断癌症和某个特定病毒之间是否有联系。如果不知道哪个事件发生在前，我们就无法知道它们之间是否存在因果关系；即使它们之间看起来好像有相关性，我们也无法知道到底哪一个是因，哪一个是果（是这种病毒引起了癌症，还是癌症让人们更容易感染这种病毒）。如果我们基于对事件发生顺序的先验观念来假定因果关系，而不是基于事件发生的真实顺序，那么就有可能误认为两者之间存在因果关系，而实际上我们所发现的不过是相关性而已。比如说，人们做了很多研究来确定肥胖和离婚这样的现象是否会因为社会关系的影响而通过社交网络传染给他人。在没有时间信息的情况下，我们无法得知这些事件的合理发生顺序。[15]

有些哲学家（比如 Hans Reichenbach）曾试图在不使用时间信息（而是从因果关系的方向性中来获知事件发生的顺序）的情况下从概率的角度界定因果关系，[16]而且有一些计算方法在特殊情况下不需要时间信息也能确定因果关系。[17]但是，绝大部分方法仍然会假定原因先于结果，并且在能够获取时间信息的情况下，使用时间信息来确定因果关系。

有些时候原因和结果似乎真的是同时发生的，所以我们无论使用时间尺度也无法区分到底哪一个在前，哪一个在后。这样的例子并不多，而其中就有一个来自于物理学的例子。在爱因斯坦－波多尔斯基－罗森（EPR）悖论中，两个粒子处于纠缠态，所以如果一个粒子的动量或坐标发生改变，另一个粒子的动量和坐标也会发生相应的变化来与之匹配。[18]这看起来似乎很反常，因为两个粒子在空间上是分离的，但这个变化却是瞬间发生的，这就让人不得不认为有的因果关系是不存在空间邻近性或时间优先性的（我们所认为的因果关系的两个关键特征）。爱因斯坦将这种异地的因果关系称为“幽灵般的超距离作用”，[19]因为超越空间的因果关

系要求信息传递的速度比光速还快，而这是不符合经典物理学理论的。[20] 这一点无论是在物理学家中还是在哲学家中都存在很多争议。[21]

有人建议使用反向因果关系（有时也叫逆向因果关系）来解决 EPR 悖论。也就是说，原因不仅可以影响将来发生的事件，还可以影响过去发生的事件。当纠缠态中的一个粒子的状态发生改变时，它会在过去的一个时间点给纠缠态中的另一个粒子发送一个信号，让另一个粒子也改变自己的状态，那么这时的状态变化就不要求信息的传递速度超过光速了（尽管这样可能会产生某种量子时间旅行）。[22] 然而在本书中，我们把"时间只能朝一个方向运动"作为一个给定的条件，把"原因必须早于结果"也作为一个给定的条件，即便我们并没有看出事件的先后顺序。

4.3　当事物随着时间变化的时候

海盗数量减少会导致全球气温上升吗？吃马苏里拉奶酪会导致人们去报考计算机专业吗？[23] 柠檬的进口数量会导致公路死亡人数减少吗？

图 4-2a 反映的是柠檬的进口数量和公路死亡人数之间的关系。该图显示，随着柠檬进口数量的增加，公路的死亡人数下降了。[24] 这些数据的皮尔逊相关系数达到了–0.98，意味着这两件事情之间存在着几近完美的负相关关系。但是，目前还没有任何人提议通过增加柠檬进口量来减少交通事故的死亡人数。

现在让我们看看图 4-2b 的情况。在这个图中，我们将进口数量和死亡人数都绘成了随着时间变化的函数。该图显示，随着时间的变化，进口数量稳定减少，而同一时期的死亡人数则在不断增加。图 4-2a 中的数据实际上也是一个时间序列——一个按照逆向时间顺序排列的序列。我们也可以用其他随着时间递减的序列（比如 IE 浏览器的市场份额、北冰洋的含冰量、美国的吸烟率等）来代替柠檬的进口数量，并从中找到完全一样的关系。

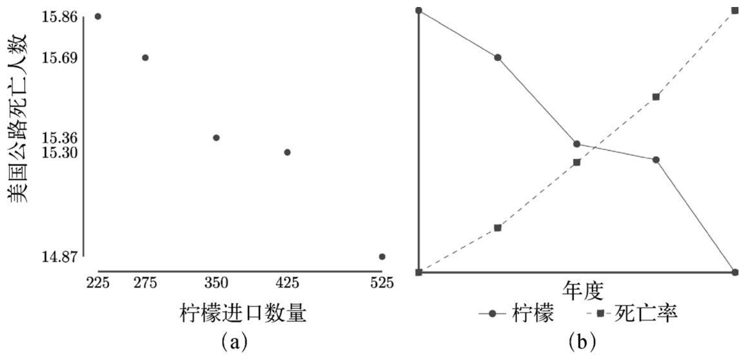


图 4-2　美国进口柠檬的数量（以公吨计算）和美国公路死亡人数（每10万人中的死亡人数）：（a）相关函数和（b）随时间变化的函数

其原因是这些时间序列是不稳定的，这意味着像平均值这样的属性会随着时间的变化而变化。即便我们改变方差来维持柠檬的平均进口数量的稳定性，但是各个年份之间的上下波动却是不稳定的。“随时间变化的电力需求不稳定”有两个原因：首先总体上来说，电力需求很可能会随着时间的变化而不断增加；其次，电力需求还具有季节性特征。而多次抛硬币的结果则是稳定的，因为每抛一次硬币，正面朝上和反面朝上的概率都是完全一样的。

出现类似的（或者完全相反的）随时间变化的趋势可能说明某些时间序列之间具有相关性，但这并不意味着它们之间就存在因果关系。相反，这正是我们寻找没有任何对应因果关系的相关性的另一种方法。如果一组股票的价格在某一段时期内都在上涨，那么即便这些股票价格每天的变化趋势都迥然不同，我们可能依然会发现它们之间存在各种相关性。在如图 4-3 所示的例子中，自闭症患者的确诊人数的增长速度似乎和星巴克咖啡店数量的增长速度相似，[25] 因为这两者的数量碰巧都是按指数级增长的，但很多其他时间序列的增长速度也是如此（比如 GDP、网页数量和

科技文献数量）。显然，这种序列中存在因果关系的可能性很小，但并不是所有序列都如此显而易见，而且很多相关的时间序列都能找到一个令人信服的解释。如果我们选择的是其他时间序列，比如装了宽带的家庭的比例，那么除了这两者碰巧都在增长以外，我们无法找到更多的证据来证明两者之间存在联系。但可能仍然有人想要找出一个解释来说明为什么这两者之间可能存在联系。然而，这依然只是一种相关性，如果我们去考察不同层次的时间粒度，或者根据数据的不稳定性做一些调整，这种相关性可能就会完全消失。

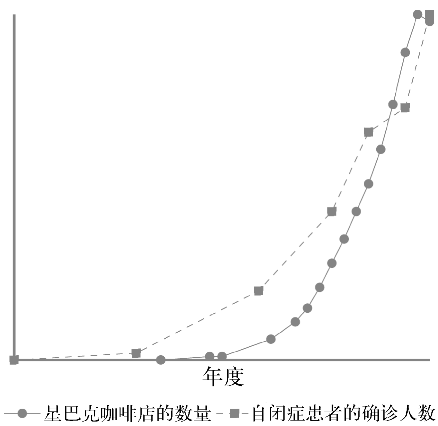


图 4-3　两个不稳定的时间序列看起来好像具有相关性，但这只是因为它们都在随着时间以指数级的速度增长

另外一种类型的不稳定性是由于被抽样调查的人群本身也在随着时间而改变。2013 年，美国心脏病协会（AHA）和美国心脏病学会（ACC）颁布了新的胆固醇治疗指导原则并发布了一个在线计算器，用于测算患者在 10 年内心脏病发作或者中风的风险。[26] 但有些研究人员发现这个计算器高估了 75%~150%的发病风险，而这可能会导致严重的过度治疗，因为用药指导原则是以每个病人发病的风险级别作为基础的。[27]

这个计算器考虑了糖尿病、高血压和当前是否抽烟等风险因素，却没有（也不可能）向患者询问所有可能会影响风险级别的因素，比如吸烟史的一些细节。等式中的相关系数（每个因素对风险级别的影响程度）是根据20世纪90年代收集的数据估算出来的,所以这项研究的隐含假定是，当前人群中的其他人口特征将会保持不变。然而，吸烟习惯和其他重要的生活因素已经随着时间发生了改变。Cook和Ridker估计，在这项纵向研究开始的时候，人口群体（白人群体）中有33%的人都吸烟，而如今同一人口群体中只有不到 20%的人吸烟，[28]这就导致风险的基线水平发生了改变，并且有可能因此导致人们过高地估计了这群人的风险级别。[29]

我们经常谈到外部效度，它指的是一个发现能否被外推到研究人群以外的人群中去（我们将在第 7 章深入探讨这个问题）。但是，还有一种效度是时间效度。外部效度指的是我们在一个地方学习到的东西如何告诉我们另一个地方将要发生的事情。在欧洲进行的一个随机对照实验的结论能否告诉我们某种药物在美国是否会有效？随着时间的变化，因果关系本身也可能会发生改变（新的规章制度会改变影响股票价格的因素），或者因果关系的强度也会发生改变（如果大多数人都在网上看新闻，那么印刷广告对人们的影响就会降低）。同样，做广告的人可能会分析出社交网络是如何影响人们的购买行为的，但如果人们使用社交网络的方式随着时间发生了改变，那么社交网络和购买行为之间的关系将不复存在（比如人们过去只会点击好朋友的主页链接，但现在会点击很多泛泛之交的主页链接）。在使用因果关系时，人们会默认那些让因果关系成立的因素是保持不变的。

如果我们考察随着时间的变化，某个医院的病人再入院率，那么就可能会出现一个与之类似的情景。从某项新政策实施之日起，或者从医院领导层变更之日起，病人的再入院率可能会随着时间的变化而增长。但是，这也可能是因为医院服务的人群也随着时间发生了改变，现在服务的人群

的健康状况比以前更差了。实际上，新政策本身可能也改变了这个人群，我们将在第 9 章详细讨论这个问题。我们常常试图根据因果关系来制定一些政策，但是政策本身可能也会改变一个人群。结果，最初的因果关系可能已经不复存在，从而导致干预措施失效，比如加州学校缩小班级规模的项目。在这个项目中，对教师需求的激增导致学校招聘了一批经验不足的教师群体。

这时也有可能会出现新的因果关系，比如出现一种新的致癌物。此外，变量的含义也可能会发生改变。语言就一直在演变，新的词汇不断涌现，而现有词汇可能有了新的用法（比如用贬义词来表达褒义）。如果我们发现政治演说内容和支持率之间存在相关性，而现在能赢得支持率的语言的含义发生了改变，那么这种相关性就不复存在了。结果，关于支持率上升的预测就会失败，而发表新演说的行为可能也不会有什么效果了。在一个比较短的时间尺度内，比如每天都有新的变化，但我们却没有考虑到这些变化，那么就有可能会出现这种情况。

有一些策略可以用来处理这些不稳定的时间序列。虽然我们可以直接忽视这种不稳定性，但还有一些更好的方法可以用来处理这种问题。比如，在数据足够多的情况下，我们可以缩短研究周期（如果时间序列的某个子集是稳定的），或者把一个不稳定的时间序列变成一个稳定的时间序列。

Elliot Sober 曾经介绍过一个具有不稳定性的例子。[30]这个例子如今被广泛使用，它讲的是威尼斯海平面和英国面包价格之间的关系。它们都随着时间的变化而上涨了，所以二者似乎具有相关性。如果使用 Sober 为这个例子提供的数据（如图 4-4a 所示，注意图中并未标出变量的单位），那么这两个变量之间的皮尔逊相关系数是 0.8204。尽管这两个时间序列都一直在增加，但这两个变量每年的具体增加量却是不断变化的，而我们真正想知道的是这些变化是如何相互关联的。最简单的方法就是观察这两个变

量的具体增加量，而不是那些原始的数值。也就是说，与上一年测量的数值相比，本年度海平面或面包价格上涨了多少？如果我们使用年度之间的变化值（如图 4-4b 所示），那么相关系数则会下降至 0.4714。

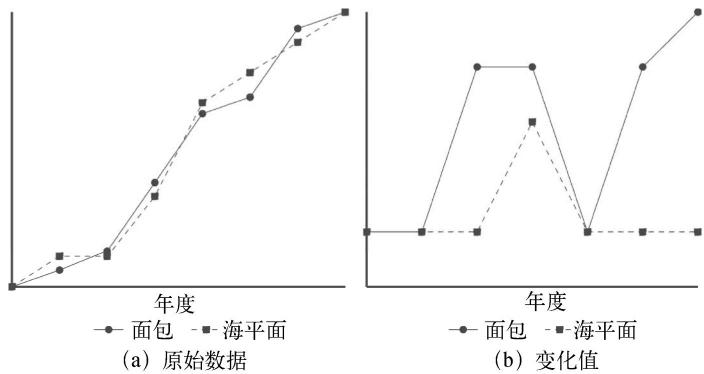


图 4-4　面包价格与海平面

这种方法叫作差分法（顾名思义，就是选取连续数据点之间的差量），它是实现时间序列稳定性的最简单方法。即便两个时间序列呈现出的长期趋势是一样的（比如一直在上涨），但如果每天或每年的变化量不同，那么二者各自的变化值可能也不再具有相关性。一般来说，仅仅采用差分法并不能保证转化出的时间序列就一定具有稳定性，要想实现时间序列的稳定性，我们可能还要采取更为复杂的数据转换措施。[31]

这就是股市数据一般使用的都是股票收益（价格变化）数据而不是真实价格数据的原因。而这正是柠檬进口数量和公路死亡人数案例问题的症结所在，也是我们在很多时间序列组中找到类似关系的原因所在。如果总体趋势相似并且具有显著性，那么这种趋势就会对相关性系数产生极大的影响，从而掩盖了短周期中两个变量的变化量（这些变化量之间可能根本不存在相关性）之间存在的差异。[32]

4.4　原因运用中的时间因素

一周中的哪一天最适合订机票？应该早上锻炼还是晚上锻炼？我们该等多久才能要求加薪？经济学家们经常谈论季节效应，这些季节效应是每年同一时间都会出现的规律，是一种不稳定性特征，但是在很多其他类型的时间序列中，比如看电影的人数（受到季节和节假日的影响）和急诊室病人的数量（可能会因为季节性疾病而剧增），我们也能发现一些与时间有关的规律。假如我们在冬季发现了一些能够让电影票销量上涨的因素，那么这些因素在夏季也许就不会起到预期的效果。还有一些规律可能只在一周中的某一天才会出现（比如由上下班的习惯导致的一些规律），或者是公共节假日的安排导致的。

事件发生的顺序可能会帮助我们把握事件发生的原因（如果我们观察到一个人先生病，然后体重才下降，那么我们就知道体重下降不可能是导致这个人生病的原因）并做出更好的预测（知道某种结果出现的时间）。但要想有效地运用原因，我们需要知道的就不只是事件发生的顺序了。首先需要知道某个关系是否只在有些情况下成立，还要知道原因和结果之间的时间间隔是多久。

因此，收集并标明时间信息至关重要。及时采取治疗措施能提高很多疾病（比如中风）的治疗效果，但是治疗效果并不总是随着治疗时间的推迟而直线下降。比如有报道称，如果我们在川崎病症状出现后的 10 天内开始治疗，将大大降低病人冠状动脉受损的风险。如果在 7 天之内开始治疗，那么效果会更好。但是，如果在 5 天之内开始治疗，治疗效果并不会进一步提高。[33] 在一些情况下，早上用药还是晚上用药也可能改变药物的治疗效果。因此，如果某种药物在临床试验中有特定的服药时间，或者每天服药的时间都是一样的，但是在非临床试验的实际使用中，每天服药的时间变化很大，那么这个药物的药效可能并没有临床试验预测得那么好。

为了确定采取行动的时间，我们还要知道一个原因需要多久才能产生某种结果。这可能意味着我们在选举之前要确定什么时候投放某些广告，在收到一条信息后要确定什么时候卖掉某个股票，或者在出行之前要确定什么时候开始服用抗疟疾药物。在有些情况下，如果我们采取的措施没有考虑到时间因素，那它们可能不会产生任何效果。比如广告投放的时间太早（后来出现的其他原因可能会干预广告效果）、股票价格还没有到达峰值就做出交易决定，或者开始服用预防性药物的时间不够早，无法起到保护作用。

时间还可能会影响我们是否采取行动的决策，因为它会影响我们对一个原因的效用和潜在风险的判断。原因的效用既取决于某个结果出现的概率（在其他条件不变的情况下，成功率为90%的原因比成功率为10%的原因要更好），又取决于出现这个结果所需的时间。人们都知道吸烟会导致肺癌和心血管疾病，但是这些疾病并不会在吸烟之后就立即出现。仅凭癌症出现的概率并不足以让我们清楚地认识到吸烟的风险，我们还需要知道时间信息。对于某些人来说，在不久的将来可能会患某种疾病的风险很小，但与在遥远的未来几乎一定会患某种疾病相比，前者的风险似乎更大。

然而，在干预措施的决策过程中，我们不仅要决定是否要采用某种措施来取得某种结果，更重要的是要决定到底采用哪一种干预措施。《宋飞正传》中有这样一个情节，Jerry 在研究各种治疗感冒的药物，他自言自语着："这个药见效快，但是这个药效长。什么时候减轻症状对我来说最重要呢，是现在还是迟些时候？"[34] 尽管这个信息增加了决策过程的复杂性，但它却能够让我们根据其他限制条件（比如一个小时后有一个重要的会议与要上一整天的课）做出更好的决策，从而规划我们的行为。

4.5　时间可能具有误导性

时间是能够让我们将因果关系从相关性中区分出来的重要特征之一。我们假定相关性存在时，只有首先出现的事件才可能是原因。然而，由于事件发生的顺序如此重要，我们反而有可能在确立因果关系的过程中过于依赖事件发生的顺序。

假设某个学校的餐厅决定减少油炸食品和高热量食品的供应量，增加水果、蔬菜和全谷物食品的供应量。自从这一措施实施后，学校学生的体重每个月都在下降。图 4-5 展示的是学生平均体重（一半学生的体重高于平均水平，一半学生的体重低于平均水平）随着时间而变化的一个虚构出来的例子。图中显示，在学校餐厅调整菜单之后，学生的平均体重骤然下降，而且这一下降趋势维持了好几个月。这是否意味着学校供应更健康的食物导致学生体重下降了呢？

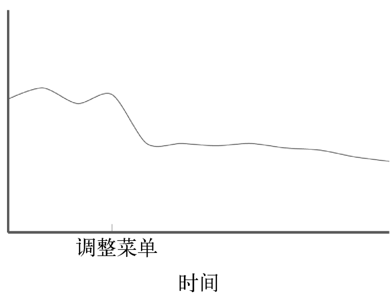


图 4-5　变量随着时间而变化的值。调整菜单之后，变量的数值下降了

某个事件发生后，一个变量的值发生了明显的变化，这样的数据常被用来证明上述类型的观点。但是，这样的数据其实并不能证明这种观点。这样的例子有很多，比如某项法律的倡导者指出，在这项法律实施后死亡率下降了；有人认为某种药物导致了一种副作用，因为这种副作用是在他开始服用这种药物的几天后出现的。

在调整菜单的那个案例中，我们并不清楚调整菜单前后在餐厅就餐的是否是同一批学生（也许那些喜欢吃健康食品的学生在调整菜单后转而开始在餐厅吃饭，而那些不喜欢新菜单的学生则不再去餐厅吃饭了），或者餐厅调整菜单是否是学生或家长的要求（因为他们正在减肥），又或者是否还有其他变化共同导致了这个结果（也许同时还增加了体育活动和休息时间）。只有一个因素发生改变，其他因素完全不变，这样的情况即便有，也极为稀少。所以，只有两个变量的时间序列会让人产生一种错觉，认为可以将某个新变量的影响完全从其他因素的影响中分离出来。尽管这是一种时间上的相关性，但也依然只是相关性而已。

现实生活中的干预措施比实验室的实验更为复杂，也更加不明确。比如说，某个区域有一家工厂，人们怀疑这个区域是癌症高发区。最终工厂被关闭，人们也采取了一些措施来恢复被污染的水和土壤。如果癌症发病率在工厂关闭后下降了，我们能否因此认为这个工厂是导致癌症高发的原因呢？实际上，我们并不清楚癌症发病率下降是否只是一个偶然事件（或者一开始的癌症高发也只是一个偶然事件），也不清楚当时是否还有其他真正导致癌症高发的因素，很多问题的答案都是未知的。此外，这些变量的值通常都很小，所以，它们的任何变化在统计学上都不具有显著性。

这就是我们熟知的一个逻辑谬论，叫作“后此，所以因此”，意思是“在此之后，因而必然由此造成”。也就是说，人们仅仅因为一件事情在另一件事情之后发生，就错误地认为后发生的事情是由先发生的事情引起的。比如说，人们可能会研究在某个特定的历史事件发生后，某些事件发生的概率是如何变化的——引入安全带法律后，交通事故的死亡率下降了吗？然而，有很多变化是同时发生的，而系统本身甚至也可能会由于干预措施的实施而发生改变（这一点我们将在第 7 章进一步讨论）。但是，也许更健康的餐厅食物只是通过促使人们增加运动量而间接导致了体重的下降。同样，如果一个体工队每次只要比赛前下雨就能打赢比赛，那么人

们有可能由此认为这两者之间存在因果关系，但其实这种现象最合理的解释就是巧合。如果我们总是盯着短期发生的事情而忽视长期的变化，那么这种问题就会经常出现。如果连续两年冬季都出现了极端的暴雪天气，那么将这两个冬季孤立起来看，我们就有可能得出错误的冬季天气规律。相反，如果我们考察了几十年的天气数据，就能在大趋势背景下了解到每年的天气波动。两个事件之所以会一起发生，可能只是因为其他因素让它们一起发生的可能性增加了。如果某种儿童疾病会在某个年龄开始出现明显特征，而这些儿童在同一年龄开始吃一些新的食物，那么很多人可能会因为这两件事情总是一起发生而认为二者之间存在某种表面联系。

还有一个相关的逻辑谬论，叫作“随此，所以因此”（与此同时发生，因而必然由此造成），它指的是在两个仅仅同时发生的事件之间找到某种因果关系。这个谬论与“后此”谬论之间的区别是：“后此”谬论涉及事件发生的先后顺序，而这也正是这种错误如此普遍的原因。

先发生的事件和最终的结果可能是由一个共同的原因导致的。（比如，治疗抑郁症的药物会让人有自杀倾向吗，还是说患抑郁症的人往往更容易自杀，也更容易服用治疗抑郁症的药物？）但这个结果的出现可能是必然的，只不过是出现在了原因之后而已。假如某个人头疼，然后吃了一些药，几个小时后头不疼了，我们是否可以说这是因为服药的缘故呢？这两个事件发生的时间让我们觉得头不疼了好像是吃药的结果，但我们无法肯定如果不吃药的话，头疼是否也会好起来。为此，我们需要做很多实验，随机选择是吃药还是不吃药，并记录下头疼减轻的速度，只有这样才能说吃药和头疼之间是否存在某种关系。第 7 章将会解释为什么这一实验还不足以证明二者之间的关系，以及为什么我们应该对比吃药的效果和吃安慰剂的效果。

两个事件在时间上的相近性可能会导致人们得出错误的因果结论。同样，因果事件之间漫长的时间间隔也可能会导致人们无法推理出二者之间的因果联系。有些结果很快就会发生（比如台球被击中后就会立即开始

运动），还有一些结果可能要经过一个缓慢的作用过程才会发生。大家都知道吸烟会导致肺癌，但从某人开始吸烟到他得肺癌之间有一个漫长的时间间隔。有些药物的副作用在服药几十年之后才会出现，锻炼会随着时间逐渐改善健康状况。但如果我们关注的是体重，那就会发现一开始锻炼时体重似乎还会上升。因为刚开始运动时，人的脂肪还未减少但肌肉却开始增多了。如果我们认为结果会紧随原因出现，那就有可能无法找到那些真正相关的因素之间的联系。从统计学上来说，科学家很难收集周期长达几十年的数据来了解影响健康的因素。对于个人来说，我们也很难将饮食和体育活动这样的因素和我们的健康联系起来。

注释

1. Leibovici（2001）。对这篇文章的评论刊载在 BMJ 杂志 2007 年 4 月 27 日出版的那一期上。
2. 这种非对称性的另一个可能的定义是：对原因进行干预会改变结果，而对结果进行干预却不会对原因产生任何影响。然而，这一定义也存在其他一些问题。因为我们通常无法对原因或结果进行干预，或者说我们在对原因或结果进行干预时，无法让其他所有变量保持不变。
3. Michotte（1946）。
4. 想要了解更多信息，参见 Joynson（1971）。
5. Michotte（1946），69，166。文中并未给出研究对象的准确描述，也未给出使用每一种描述的人数。
6. Michotte（1946），63。
7. 在早期的研究中，Heider 和 Simmel（1944）使用更加复杂的动作制作了一个相似并且更长的视频。在没有任何提示的情况下，所有参与者都使用描述生命体的方式来描述视频中发生的事件。虽然视频中的物体只是一些三角形和圆形，但参与者却将它们描述成了具有不同意图的生命体，而且这些生命体正在进行各种活动，比如战斗和追逐。
8. Michotte（1946），249，347。
9. 在 Beasley（1968）的研究中，有 64% 的参与者认为这些动作之间是有因果关系的，而在 Gemelli 和 Cappellini（1958）的研究中，有 87%的参与者认为这些动作之间是有因果关系的。

10. Michotte（1946），347。
11. Buehner 和 May（2004）。
12. Greville 和 Buehner（2010）；Lagnado 和 Speekenbrink（2010）。
13. Faro 等（2013）。
14. Bechlivanidis 和 Lagnado（2013）。
15. 友谊往往发生在那些具有很多共同特征（相似的个性或共同的环境）的人之间。由于这些共同特征的混淆效应（通常观察不到的），即使我们掌握了时间信息，通常也无法将这些解释区别开来。
16. Reichenbach（1956）。
17. 想要了解更多关于贝叶斯网络的信息，参见 Scheines（1997）。
18. Einstein 等（1935）。
19. Born 和 Einstein（1971）。
20. 虽然 EPR 悖论最初是作为思想实验被提出的，但后来 Ou 等（1992）通过实验证实了这一悖论。
21. 想要大致了解这一内容，参见 Cushing（1998）。
22. 想要了解更多关于时间及时间旅行的信息，参见 Price（1997）和 Lewis（1976）。
23. tyler vigen 网站能够自动在不同的时间序列之间生成各种相关性。
24. Johnson（2008）最早使用了这个案例。关于死亡人数的数据来自 FARS Encyclopedia 网站。关于柠檬的数据是从 Johnson（2008）的原始数据中估算出来的。
25. 这些数据来自 Autism Speaks 网站和 Telegraph 网站。
26. Stone 等（2013）。
27. Ridker 和 Cook（2013）。
28. 数据来自美国疾病控制与预防中心网站。
29. 对该计算器的这一批判引起了人们的争议，有些人暗示对照组报告的中风和心脏病突发事件的数量要少于实际发生的数量，参见 Muntner 等（2014）。
30. Sober（1987，2001）。
31. 人们可以重复地将数据进行区分，也可以将不同年份的数据进行区分，以便消除季节性的影响。想要了解关于稳定性的经典测试，参见 Dickey 和 Fuller（1981）；Kwiatkowski 等（1992）。
32. 想要了解人们针对差异法提出的反对意见，参见 Reiss（2007）。
33. Newburger 等（2004）。
34. David 等（1991）。

第 5 章　观察法

如何仅通过观察事物的运行方式来把握事件发生的原因？

有一天下班，我在地铁上看到一则广告。广告上这样写道："如果你高中毕业了，找到了一份工作，并且婚后才生的孩子，那么你 98%不会穷困潦倒。"这则广告的目的是呼吁十几岁的女孩不要早孕，但我们并不清楚该如何理解这个统计数据。这句话的意思似乎是，如果一个年轻女孩能够按照广告上说的那样做，那么她 98%不会穷困潦倒。但是，事实真的是这样吗？而且这句话是说她现在不会处于穷困潦倒的境地，还是永远都不会处于穷困潦倒的境地呢？这个数据是从一项研究中得出的，这项研究考察了不同婚姻状况、年龄和教育水平等特征的人口，计算了总的贫困人口比例以及各个人群中贫困人口所占的比例。[1] 但是，统计结果完全建立在观察到的数据的基础之上。

没有任何（个人或社会的）政策能强制年轻女孩怀孕或者不怀孕，也没有任何政策能迫使她们穷困或者不穷困。这就意味着这个数据只统计了我们观察到的一部分人口中的一个特征：在我们观察到的高中毕业、找到工作并且婚后才生孩子的人口中，有 98%的人并未穷困潦倒。但如果具体到某一个人，她高中毕业，找到了一份工作并且婚后才生的孩子，那么她贫困潦倒的概率可能和统计数据并不一样。这一点类似于第 1 章讨论的

SIDS 案例。在那个案例中，我们发现任意一个家庭的孩子患上 SIDS 的概率和具体某个家庭的孩子患上 SIDS 的概率是不一样的。

而且，有些人没有完成学业的原因可能也正是导致他们贫困潦倒的原因，并且这些原因是他们不可控的。也许他们不得不照看家里的老人，也许他们缺少生活保障（比如医疗保障）或家人的支持。这就意味着他们可能无法只是简单地去寻找一份工作，而且不得不去解决其他问题（比如为父亲或母亲另找一个护工）。而且，如果这些其他因素（比如高额的医疗费用）才是最终导致贫穷的原因，那么即便满足了上述三个标准，他们陷入贫困境地的风险也不会改变。如果未完成学业、找不到工作和婚前生子只是那个导致人们陷入贫困境地的因素所带来的其他影响，那么针对这些问题采取干预措施就像在处理事情的结果而不是起因。贫穷可能是情境因素引起的，而且这些情境因素是很难干预的，比如歧视、工作机会匮乏或者教育水平低下等。

这对公共政策的制定有着巨大的影响。如果我们只致力于提升人们受教育的机会和就业机会，却不知道是什么因素导致人们无法获得这两个机会，也不知道这两者本身是否就是导致贫困的原因，那么我们就更难制定有效的干预措施了。我们不知道是否还有其他问题导致我们无法实现经济保障，也不知道我们所采取的措施能否给我们带来想要的结果。此外，所有这些因素都有可能是贫穷导致的结果，我们也许应该通过新的方式来直接解决贫穷问题。[2]我们将在第 7 章和第 9 章详细介绍如何采取干预措施才能取得预想的效果，以及我们需要哪些信息才能预测出某个干预措施的效果。

如果我们能够强迫一些人读完高中（或者不读完），然后将他们随机分配到这些不同的实验组中（避免他们出现其他情况），就有可能将这一行为对未来经济形势的影响分离出来。但实际情况是，我们所观察到的数据往往是我们所能获得的全部信息。如果为了考察年轻女孩怀孕是否是贫

穷导致的结果或引起贫穷的原因（或者是否存在一个反馈循环）而去做一些实验，那么这种行为是不道德的。研究人员还需要确定接触某些媒体对人们的影响，比如某个竞选广告是否影响了公众的舆论？电视剧《十六岁的怀孕女孩》是否影响了年轻女孩的怀孕比例？在这种情况下，我们无法控制人们接触媒体的行为，甚至无法确定某个人是否接触过某个媒体。研究人员通常只能依靠媒体市场的总体特征——在某个地区投放某个广告后，过一段时间这个地区的民意测验结果与其他地区相比发生了什么变化。我们可能无法在一个很长的时间周期内追踪参与者的行为，即便可以追踪，那些实验费用可能也贵得吓人。弗雷明汉心脏研究[3]在长达几十年的时间周期里坚持不懈地跟踪调查一个巨大样本群体，这需要投入巨大的研究精力，但这是研究活动中的特例而不是惯例。

本章讨论的内容是，当我们只能观察正在发生的情况时，如何去发现事物的运行方式。我们还将讨论这些方法的局限性，以及观察数据通常存在的一些局限性。

5.1 规律性

5.1.1 穆勒五法

假设一群计算机科学家参加了一个编程马拉松。这些科学家们每天都忙到凌晨，营养均衡和饮食健康对他们来说简直就是天方夜谭，所以很多人在熬夜时都是依靠浓咖啡、比萨饼和功能性饮料补充能量的。不幸的是，在第二天的颁奖典礼上，他们中的很多人都生病或者缺席了。我们怎样才能确定是哪些因素导致他们生病的呢？

有些团体中出现了某种结果，而另一些团体中没有出现某种结果，针对这种情况，试图找出这些团体中的共同点和不同点，这是 John Stuart Mill

在 19 世纪提出的穆勒五法的典型用途之一（其中涉及食物中毒的案例似乎还挺多）。[4]

首先可以想一想：所有出现某种结果的案例之间有什么共同点？如果在所有头疼的案例中，唯一的共同点就是人们都喝了功能性饮料，那么这就在一定程度上证明了功能性饮料可能会导致头疼。这就是穆勒所说的契合法。在表 5-1 所示的例子中，我们只对头疼的案例感兴趣，所以只看表中出现头疼症状的那几行数据。先来看看哪些案例中出现了头疼的症状，然后再来看这些案例都有什么共同特征。我们注意到，这些案例唯一的共同之处在于他们都喝了功能性饮料，所以根据契合法，功能性饮料就是导致他们出现头疼的原因。

契合是指某个原因是导致某种结果的必要条件——除非出现这个原因，否则不会出现这种结果。然而，这并不意味着这个原因**每次**都会导致这种结果。如果那样的话，这个原因就成了出现这种结果的充分条件。[5] 在表 5-1 中，Betty 也喝了功能性饮料，但是她却没有出现头疼的症状。因此，我们不能说喝功能性饮料是出现头疼症状的充分条件，只能说这些条件对于我们观察到的内容来说是真实的。我们永远无法从有限的样本中去证实必要条件或充分条件。

表 5-1　根据穆勒的契合法，我们发现喝功能性饮料会导致头疼

	咖啡	比萨饼	熬夜	功能性饮料	头疼
Alan	X	X	X	X	是
Betty	X		X	X	否
Carl		X		X	是
Diane			X	X	是

这种方法有一个局限性：它要求每一个案例都是一致的。如果有几百个人都生病了，只有一个人没生病，那么我们也无法找出某种因果关系。值得注意的是，这个方法没有考虑到 Betty 也喝了功能性饮料却没有出现

头疼的情况。这就是这个方法只能让我们找到必要条件而不能找到充分条件的原因——它没有包含出现了某个原因却没有出现某种结果的情况。

要想确定充分条件，我们就要考察出现某种结果和未出现某种结果的情况有什么差别。如果所有熬夜的人第二天都很疲惫，而那几个没有熬夜的人第二天都很精神，那么我们就会发现熬夜是第二天很疲惫的一个充分条件（在本例中）。这就是穆勒的差异法。

在表 5-2 中，我们对比了疲劳案例和非疲劳案例之间的差异。注意，在所有疲劳案例中，四个因素的情况都是一样的，所以我们无法使用契合法来确定其中一个因素就是导致疲劳的原因。通过考察这些案例的差异，我们看到熬夜似乎是出现疲劳状况和未出现疲劳状况的唯一差异。与契合法一样，这个条件相当严格。因为有可能会碰巧出现一些情况：即便熬夜仍然是导致疲劳的原因，但其他因素的情况却可能不尽相同。我们将在下一节介绍概率法，这种方法使用的是事件出现的相对频率，它对关系的要求没有这么严格。

表 5-2　通过穆勒的差异法，我们发现熬夜会导致疲劳

	咖啡	比萨饼	熬夜	功能性饮料	疲劳
Ethan	X	X	X	X	是
Fran	X	X	X	X	是
Greg	X	X		X	否
Hank	X	X	X	X	是

概括一下，如果没有某个原因，某个结果就不会出现（即每次出现这个结果之前都会出现这个原因），那么这个原因就是这个结果的必要条件；如果每次只要出现某个原因，就一定会出现某个结果（每次某个原因出现之后必然伴随某个结果），那么这个原因就是这个结果的充分条件。某个原因可能是必要条件但不是充分条件，反之亦然。在编程马拉松的案例中，每次出现疲劳状况时，之前必然熬夜了，这样熬夜就成了出现疲劳状

况的必要条件。但是，这并没有说明熬夜就是出现疲劳状况的充分条件（有可能有些人熬了夜却不疲劳）。同样，每次喝了功能性饮料之后都会出现头疼症状，这就说明喝功能性饮料是出现头疼症状的充分条件，却没有说明喝功能性饮料是否是头疼的必要条件（因为可能还存在其他导致头疼的因素）。

还有一些原因可能是某个结果的充分必要条件（如表 5-3 所示）。为了找出那些既是充分条件也是必要条件的原因，我们将契合法和差异法结合起来使用，这就是穆勒的契合差异并用法。在这种情况下，我们要找的是那些每次出现某种结果时都会出现的因素，并且只有在出现这种结果时才会出现这些因素。在表 5-3 所示的例子中，两个肚子疼的人都熬夜了，也都喝了咖啡。所以根据契合法，这两个因素可能是导致肚子疼的原因。现在，我们再来考察一下这两个因素在那些肚子不疼的人和肚子疼的人身上有什么差别。我们发现，Diane 熬夜了，但没有出现肚子疼的症状。所以熬夜并不满足差异法的要求，而喝咖啡却满足这一要求，因为所有喝了很多咖啡的人都出现了肚子疼的症状，而没有喝咖啡的人都没有出现这一症状。因此，根据这个表格的数据来看，喝咖啡是肚子疼的充分必要条件。

表 5-3　根据穆勒的契合差异并用法，我们发现喝咖啡是导致肚子疼的原因

	咖啡	比萨饼	熬夜	功能性饮料	肚子疼
Alan	X	X	X	X	是
Betty	X		X		是
Carl		X		X	否
Diane			X	X	否

这个方法存在的问题是什么呢？假设我们看到有 2000 人在吃了没清洗的水果后生了病，但是还有 2 个人吃了水果后居然没有生病，还有几个人在吃了没有烤熟的鸡肉后出现了食物中毒。按照穆勒的方法，我们无法

在吃水果和生病之间找到任何因果联系，因为吃了没清洗的水果既不是生病的必要条件也不是充分条件。在现实生活中，有很多因果关系并不是每次都会出现的，所以穆勒要求的条件过于严格了。一般而言，我们不能仅凭几个反例就完全推翻一个原因，但这种方法仍然能为我们提供一个直觉式的指导原则，帮助我们探索各种因果假设，而且这种方法与我们对原因的一些定性研究所用的方法是一致的。[6]

在现实中，只有一个原因和一个结果的情况也很少见。也许人们吃比萨饼、熬夜并且喝了大量咖啡，结果导致同时出现很多疾病。如果我们看到人们既出现了疲劳症状，又出现了肚子疼的症状，但是这些既疲劳又肚子疼的人之间并没有什么共同点，或者这些人和其他人并没有什么差别，那我们应该怎么做呢？有些情况下，我们可以将导致疲劳和肚子疼的原因区分开来。

在表 5-4 所示的例子中，假设我们已经知道熬夜是导致疲劳的原因。这样一来，就可以用熬夜来解释 Alan、Betty 和 Diane 感到疲劳的事实了。由于我们已经知道熬夜并不会导致肚子疼，所以可以假设一定有其他因素导致了肚子疼。然后，我们只要考察一下所有肚子疼的案例之间有什么共同点和不同点就可以了。一旦排除了疲劳和熬夜这两个因素，剩下的唯一一个共同因素就是喝咖啡了。尽管熬夜也是那些肚子疼的人所共有的特征，但穆勒假定我们基本上可以排除那些已知的因果关系。如果我们知道熬夜会导致疲劳，那么就可以考察在这个原因和结果被排除后还剩下什么。如果还剩下一个原因和一个结果，那么这个原因就一定是导致剩下的这个结果的原因。这就叫作剩余法。不过，这个方法假定了我们已经知道所有其他可能的原因导致的所有结果，并且一个原因只会导致一个结果。如果现实情况是，熬夜和喝咖啡相互作用才会导致肚子疼，那么我们就无法通过剩余法来找到这种因果关系了。

表 5-4　根据穆勒的剩余法，我们发现喝咖啡会导致肚子疼

	咖啡	比萨饼	熬夜	功能性饮料	疲劳	肚子疼
Alan	X	X	X	X	是	是
Betty	X		X		是	是
Carl		X		X	否	否
Diane			X	X	是	否

我们可以根据这个方法做出一些假设，以此来推断可能是什么原因引起了我们观察到的现象，但我们无法用它来证实某个关系是因果关系。接着来看看变量集或者变量的来源。我们研究的变量永远都只是所有可能被衡量出的变量的一个子集，它们可能只是我们根据感知到的相关性选择出来的，也可能只是我们在事后分析数据时实际衡量出的变量。

然而，真正的原因可能并不在那些假设之中，而这可能会让我们无法找到导致某种结果的原因，或者找到的可能只是表示原因的一个迹象。如果每个吃比萨饼的人同时也喝了一些有问题的自来水，而我们的变量集中却没有包含喝水这个变量，那我们将会发现吃比萨饼是导致某种结果的一个原因。尽管它其实并不是一个原因，但它却与喝水这个变量有一定的关系。在这个案例中，即便我们将喝水这个变量考虑进去，如果吃比萨饼与喝水这两个因素总是共同出现（每个吃比萨饼的人都喝水，每个喝水的人都吃比萨饼），我们还是无法确定吃比萨饼就是导致某种结果的原因。但其实在这种情况下，喝水和吃比萨饼这两个因素似乎都会成为导致某种结果的原因。因为我们从来没有分别观察过这两个变量，所以只能看到这两个潜在的原因和结果之间存在一种完美的规律性。这个问题不是穆勒的方法独有的，而是我们根据观察数据寻找因果关系时的一个比较普遍的问题。然而，如果我们能够使用实验法，让参与者在吃比萨饼的时候不喝水，或者喝水的时候不吃比萨饼，就能解决这个问题了。然后我们就会看到只有那些喝了水的人（无论他们吃不吃比萨饼都一样）才会生病。

再来看看计算机科学家的那个案例。也许那些程序员在工作的时候很容易吃过多的比萨饼。如果吃过多的比萨饼会导致体重增加，那么随着比萨饼食用量的增加，我们应该能够看到这些人的体重也随之增加。这就是穆勒的共变法。在共变法中，原因和结果之间存在剂量效应——随着原因剂量的增加，结果的剂量也随之增加。如果一项研究声称咖啡可以降低某个年龄段之前的人群的死亡风险，那么我们就会认为每个人的咖啡饮用量会影响他们的死亡风险。而如果一天喝 1 杯咖啡和一天喝 10 杯咖啡的结果完全一样，那么实际降低死亡风险的很有可能是与喝咖啡共同出现的其他因素。

当然，实际情况总是更复杂一些，原因和结果之间的关系也并不总是线性关系。举个饮酒的例子：随着饮酒量的增加（在一定范围内），饮酒对健康的好处也会增加，但如果饮酒过量就会成为一种非常不健康的行为。有一个 J 形曲线反映了饮酒量与冠心病等健康问题之间的关系（如图 5-1 所示）。在每日饮酒量从 0 克上升到 20 克（大约两小杯）的过程中，疾病的发病率逐渐降低，但到了 20 克以后，随着饮酒量的增加，发病率开始上升。[7] 与之类似的关系还有，人们假定运动强度和感染疾病的概率之间存在的联系，[8] 以及喝咖啡和心力衰竭等健康问题之间存在的联系。[9] 和吃药一样，饮酒、喝咖啡以及运动等活动都有一个量，超过这个量之后，它们就可能会危害健康。因此，过了某个量之后，我们就看不到剂量效应了；相反，我们会看到该因素产生的效果开始逐渐减弱而不是不断增强。

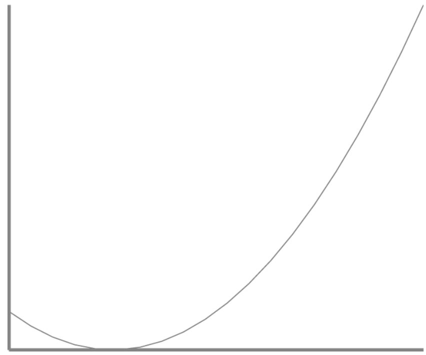


图 5-1　J 形曲线

John Snow 发现了 1854 年伦敦爆发霍乱的原因，这是历史上使用穆勒五法的最有名的案例之一。[10] Snow 并没有明确宣称自己采用了穆勒的方法，但他所用的研究方法和穆勒的方法是建立在相同的原理之上的。疫情暴发时，人们并不清楚疾病究竟是如何传播的，但有一张地图显示疾病的发病率在不同地域间存在明显差异。这个疾病是人传染给人的吗？居住地有什么东西引起了疾病的暴发呢？还是说因为人们都生活在被感染的区域所以才导致了疾病的暴发？

Snow 发现，很多死亡都发生在某个特定的地理区域，而且还都靠近宽街（Broad Street）的水泵：有一些房子更加靠近另外一个街泵，而这些房子里的居民只有 10 个人死于霍乱。在这 10 个人中，有 5 个人的家属告诉我们，这些死者总是去宽街的水泵取水，因为他们更喜欢喝那里的水。还有另外 3 个死者都是小孩，他们上学的地方靠近宽街的水泵。[11]

Snow 发现，死者大多数都可能使用过宽街水泵里的水。他又考察了那些似乎和大多数死亡案例不一致的案例，发现这些人虽然不住在宽街水泵附近，但使用的也是宽街水泵里的水。这里用的正是穆勒的契合法——在所有出现某种结果（比如感染霍乱）的案例中寻找共同点。Snow 还使用

了差异法，因为他写过："在伦敦的这个区域，除了那些习惯饮用上述泵井中的水的人，还没有出现其他感染霍乱的特殊情况。"[12]他证实了霍乱的发病率在使用了那个水泵的人群中上升了，而且也只在那个人群中上升了。

5.1.2 各种复杂的原因

穆勒五法有一个问题：一个原因导致某种结果出现的可能性的大小，取决于除这个原因之外还存在哪些其他因素。比如说，分别服用两种药物可能对血糖没有任何影响，但同时服用可能就会产生相互作用，从而显著地提高血糖值。要想解决这个问题，就不能只看单一原因和单一结果之间成对的关系，而是要考虑导致某种结果的一系列因素的组合。比如说，这一起交通事故可能是酒驾和汽车间距太近共同导致的，而另一起交通事故可能是能见度太低、路面结冰和鲁莽驾驶共同导致的，还有一起交通事故则可能是发短信和超速共同导致的。

我们知道，在流行病学研究中，各种原因都是相互关联的，人们与各种环境的长期接触、生活方式以及严重暴露（比如接触某种传染性疾病）等各种因素会共同影响人们的健康状况。这种情况在流行病学领域经常出现。因此，流行病学家 Kenneth Rothman 提出用饼形图来表示这些原因组合。[13]原因饼形图是由一系列足以导致某种结果的因素共同组成的，它包含能产生某种结果的所有必要因素。图 5-2 展示的是导致三起交通事故的原因组合。

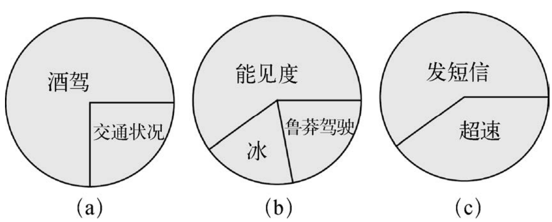


图 5-2　三起交通事故的原因组合

在这个例子中，每一个饼形图都足以让这个结果发生，所以每一次这些因素出现时都会发生一起交通事故。然而，由于有很多不同的原因组合都能导致这个结果，所以这些因素中的每一个因素都不是引起交通事故的必要条件。休谟和穆勒的要求是，每次某个原因出现时都会导致相应的结果，但有时候让这个原因起作用的那些必要条件可能根本就没有出现，或者每次只有在出现某个原因时才会出现相应的结果。但有时不同的原因能导致同样的结果。因此，休谟和穆勒的要求是一个极为严格的条件。在现实生活中，很多结果可能是通过多种方式产生的，并且这种情况往往都存在一系列原因。

然后，原因的概念就变成了因素组合中的一个组成部分，这组因素在一起时足以导致某种结果，但这个组合可能并不是出现某种结果的必要条件，因为像这样的原因组合可能有很多。这就是 John Leslie Mackie 的方法，他认为原因不过是那些 INUS 条件（非必要充分条件中的非充分必要部分）而已。[14]在饼形图那个例子中，每一块单独的饼形图都是不充分的，因为要想产生某种结果还需要和其他几块结合起来共同起作用。但是，每一块都是必要的，因为如果缺少其中任何一块，这种结果就不会出现。另一方面，每个饼形图本身又都是非必要条件，因为可能会存在多个饼形图，它们中的每一个都足以导致这样的结果。因此，我们不应该只将经济因素、其他政党的广告或者支持率锁定为导致某个竞选结果的主要原因，而是要再现所有的影响因素，或许还可以尝试分析它们各自的重要性。

然而，并不是所有的原因都是 INUS 条件。比如说，因果关系可能并不是永远不变的，所以即便我们拥有所有可能的信息，而且所有的必要条件也都出现了，但是结果却并不会百分之百出现。这叫作非决定论，放射性衰变就是一个例子。在这个过程中，我们永远无法确定某个粒子是否会在某个具体的时间发生衰变，只能知道这件事情发生的概率。衰变永远都不会有 INUS 条件，因为衰变是没有充分条件的。同样，如果变量组合选择

不当（比如比萨饼和水的例子），也可能会出现一些并不是构成原因的、表面上的 INUS 条件。这些推理的准确性和完整性永远取决于数据的完整性。

5.2 概率

5.2.1 为什么要使用概率

本章是由一则广告上的一句话展开的，那句话是这样的："如果你高中毕业了，找到了一份工作，并且婚后才生的孩子，那么你 98%不会穷困潦倒。"这句话试图暗示一个因果关系：对于一个人来说，如果高中毕业、找到工作和婚后生子这些条件都成立，那么她就有 98%的概率不会贫穷。人们之所以会对这个统计数据如此感兴趣，是因为这个概率十分接近 100%。但是，如此高的概率并不意味着这个关系就是因果关系。可能有些关系出现的概率很高，但并不是因果关系，还有一些关系是真正的因果关系，但这些因果关系中的原因可能只是降低了结果出现的概率，或者并未改变结果出现的概率。那么因果概率的理念到底有什么用处呢？

我们之所以需要使用概率法（这个方法并不要求原因能百分之百导致相应的结果，也不要求每次出现某个结果之前都会出现某个原因），是因为有些关系本身就是不确定的，比如放射性衰变的例子。在这些情况下，即便穷尽毕生所学，我们仍然无法确定某个结果是否会发生。因为这种情况既不存在前面几种方法要求的规律性，也不存在具有任何规律的变量组合。物理学领域经常出现具有不确定性的案例（比如量子力学），但这种案例在日常生活中更加常见，比如设备出现故障的时候。

在很多情况下，我们之所以认为有些事物看起来具有不确定性，只是因为我们缺乏对事物的认知——即使这些事物的所有信息都是完全可以预测出来的。并不是每一个接触石棉的人都会得癌症，有些药物只会在

一部分病人身上出现副作用，而且某些类似的情况也并不是每次都会导致股市泡沫。如果掌握了某种药物的作用机制，或者能够观察到足够多这种药物的副作用案例并且知道它在哪些人身上产生了副作用，那我们就能找到这种药物产生副作用的必要条件。

绝大多数情况下，我们需要处理的不仅是观察数据（我们不能强迫人们去吸烟以便观察谁会得癌症），还要处理不完整的数据。这可能意味着我们正在错过一些变量（有氧代谢能力可能是估算出来的，而不是通过在跑步机上进行最大摄氧量测试测量出来的），并且只能观察到一个有限的时间段（手术结束一年后的恢复状况，而不是三十年后的恢复状况），还可能意味着样本之间的时间间隔比我们想要的大得多（每小时的脑代谢情况，而不是像脑电图一样的数据）。这可能是为了节约成本（对于大规模的研究来说，最大摄氧量测试不仅成本高、耗时长，而且对一些身体不够健康的参与者来说可能也不安全），也可能是数据采集的可行性导致的（我们几乎不可能花费几十年的时间去追踪研究某个参与者），还可能是技术上的局限性导致的（用微透析技术来测量代谢活动是一个缓慢的过程）。在使用概率法时，很容易混淆这两种适用于不同原因的概率，一个适用于缺乏认知的情况，一个适用于不确定的关系本身。但是，我们一定要记住这是两种不同的概率。

之所以要使用概率来定义因果关系，是因为我们不仅想要知道某个事物到底是不是原因，还想要知道这个事物到底有多重要。具体来说，就是我们想要将某种药物的常见副作用和罕见副作用区分开来，或者想要找到最有可能增加就业机会的政策。要想量化某个原因对某个结果造成的影响，可以在使用连续变量的情况下测量这个结果的大小（比如在某些新闻播出之后，某个股票的价格上涨了多少），或者在使用不连续变量的情况下测量这个结果发生的概率（比如某个股票价格上涨的可能性有多大）。

然而一般情况下我们看到的关于因果关系的报道只会表明某个原因

增加了出现某种结果的风险。下面几行文字摘自几篇关于科技论文的报道的开头部分。

> “科学家们做过报告，说治好抑郁症患者的失眠问题可以将他们完全康复的概率提高一倍。”[15]
>
> “哈佛大学公共卫生学院研究人员的一项最新研究显示，每天喝几杯咖啡似乎能让男性和女性自杀的风险减少 50%左右。”[16]
>
> “科学家们在本周三的报告中讲到，随着年龄的增长，随机突变的基因数量越来越多。因此，大龄男子比年轻男子更有可能拥有一个患有孤独症或精神分裂的儿子。这是科学家们第一次对这种逐年增加的影响进行量化研究。”[17]

有很多关于科技论文的报道在开头只会提一下风险的增加或减少，而增加或减少的准确数据则要留到几段之后才说。即便如此，这些例子中给出的信息仍然是相对的：概率增加了一倍或风险降低了一半。对一些事件来说，概率增加了一倍可能听起来差别很大，但如果这只是将一件事变成了两件事，那这个概率的说服力就大打折扣了。比如说，中风的风险可能从 0.0000001 增加到了 0.0000002，也可能从 0.1 增加到了 0.2。这两种情况的概率都增加了一倍，但在第一种情况下，增加一倍的是一个很小的数字，而增加后的数值仍然是一个很小的数字。图 5-3 向我们直观地展示了这一差别。在第一种情况下，一千万个事件中只有一个这样的事件，增加一倍之后也只有两个这样的事件。图中每一个黑点表示一个这样的事件，而每一个灰点表示一万个事件。因此，即便相对风险都增加了一倍，但在了解这一概率的绝对值后，我们可能就会改变原本的观点。后面在介绍实验的实施与评估以及做决策等内容时，要牢记这种结果大小或概率增加问题。当你在阅读关于新的科学发现的报道时，也要想想这个问题。

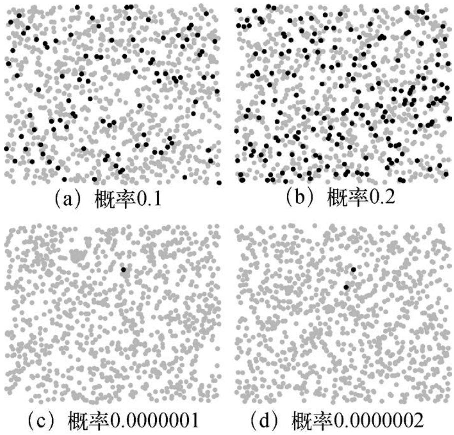


图 5-3　在上面两张图中，每一个点代表一万个事件。而在下面两张图中，黑点代表的是一个事件，灰点代表一万个事件。从左图到右图，黑点所代表事件发生的概率都增加了一倍，但也要考虑事件的总数

考察样本的大小（比如研究的人口群体有多大）尤为重要，如果观察的样本数量不显著，我们可能都无法将那些结果区分出来。[18]某个差异的出现可能仅仅是自然变化、噪声或测量失误引起的。比如说，根据风险因素的不同，蛛网膜下腔出血（SAH，一种极为罕见却致命的中风）的症状每年在 10 万人中只会出现 8 例。[19]这就意味着，如果我们用一年的时间去观察 10 万人，或者用 10 年的时间去观察 1 万人，只能看到 8 起这种中风事件。而如果观察一个更小的样本，那么我们观察到的这一事件发生的概率就会远远低于它真正发生的概率。因为在一个小样本中，我们可能会观察到 8 起这样的事件，也可能会观察到 0 起这样的事件，这就会导致我们对这种风险得出错误的结论。

5.2.2　从概率到原因

休谟的研究方法的核心是原因和结果之间存在的规律性，而概率法

的基本理念则是原因让结果出现的可能性更大。

如果一件事与另一件事之间没有因果联系，那么在第一件事出现后，第二件事出现的概率应该不会发生任何变化。抛硬币时正面朝上和反面朝上的概率都是 50%，而且每一次抛硬币都是一个独立的事件，所以每一次正面或反面朝上的概率并不会因为上一次抛硬币的结果而发生改变。即使上一次抛硬币的结果是反面朝上，那么接下来每一次抛硬币时正面朝上的概率依然是 50%。图 5-4a 用分布图（也叫镶嵌图或矩阵图）展示了这种情况。横轴代表的是第一次抛硬币时可能出现的结果（正面或反面），纵轴代表的是第二次抛硬币时可能出现的结果（也是正面或反面）。长条的宽度代表的是第一次抛硬币时正面或反面朝上的概率（如果这个硬币十分不公正，那么第一个条形可能会十分狭窄），而灰色长条的高度代表的是第二次抛硬币时正面朝上的概率。（剩余区域代表的是第二次抛硬币时反面朝上的概率。）因为每一种结果出现的概率都是完全一样的，所以图中每个部分的大小都是一样的。[20] 相反，由于政治信仰和政治联盟的不同，某个人成为副总统候选人的概率则会因为总统候选人的不同而上升或下降，所以这些事件是相互依赖的。

从直觉上来说，如果某件事会导致某个结果，那么在这件事发生后，这个结果出现的可能性应该比平时高得多。由于蚊子会传播疟疾，因此如果某个地区的蚊子感染了疟疾，那么这个地区的疟疾发病率应该更高。原因也可能会让某个事件发生的可能性变小，或者说原因也可能会让某个事件不发生的可能性变大。如果钾能够减缓肌肉痉挛现象，那么人们在服用了钾之后，肌肉痉挛的病例就应该有所减少（如图 5-4b 所示）。在这个图中，服用钾的概率比未服用钾的概率要低，所以我们用一个更窄的长条来表示服用钾的概率。然而，这个长条形的大部分都是阴影，因为在服用了钾之后，未出现痉挛症状的概率要高于出现痉挛症状的概率。相反，在未服用钾的情况下，出现痉挛症状的概率比未出现痉挛症状的概率要高得多。

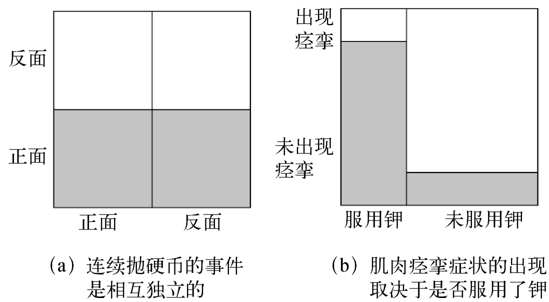


图 5-4　这两个图表示的是条件性概率。一旦你选定一个沿着底部横线发生的事件（比如服用钾），那么另一个事件（比如出现痉挛症状）的概率就可以通过长条的阴影部分来表示。在服用钾（小条形）之后，出现痉挛症状的可能性下降。但是，每一次抛硬币（同样尺寸的条形）之后，下一次抛硬币时正面朝上或者反面朝上的概率都相同

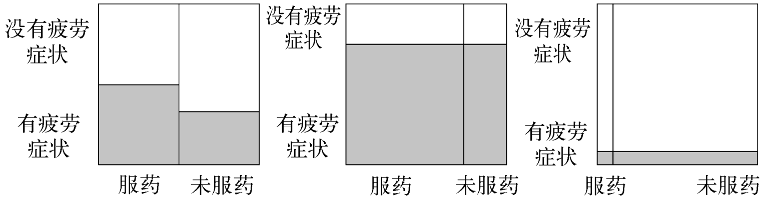
概率的提升或降低可能导致人们错误地将没有因果关系的两个事件联系在一起（非原因的事件似乎也能提升某种结果出现的概率），还可能导致人们错过两个事件之间的因果关系（不是每一个原因都能提升某种结果出现的概率）。我们在第 3 章初步探讨了相关性以及相关性产生的方式。相关性的产生有时完全是因为巧合；有时可能是因为我们验证的假设太多，所以必然会碰巧发现一些似乎具备显著性的关系；也有可能是因为所用的变量并不能准确反映真实的原因。有人说某个节食方案能在一定程度上起到减肥的作用，但起到减肥效果的相关变量可能只是节食的行为，而不是我们正在验证的这个节食方案。还有一种可能，虽然我们只考察了两个因素之间的一种关系，但可能会由于一些结构因素而发现很多相似的关系。第 3 章还讲了人们是如何发现一个国家的巧克力消费量和该国获得诺贝尔奖的人数之间存在相关性的。也许红酒、奶酪或咖啡的消费量和获得诺贝尔奖的人数之间也同样有着很高的相关性。有一项研究发现，除了其他因素之外，诺贝尔获奖人数和宜家家居（IKEA）的店铺数也有相关性。[21] 因此，巧克力的消费量可能只是一个反映某种人口特征的指标，

这个特征可能是一个国家的财富和资源。而这个特征才是既让人们消费更多巧克力，又提升人们获得诺贝尔奖概率的原因。

有时，一个变量似乎会提升另一个变量出现的可能性，但并不是真正导致另一个变量出现的原因。这种现象往往是由上面这种共同原因导致的。比如说，经济衰退既能导致通胀率下降，也能导致失业率上升，而通胀率下降和失业率上升这两个因素好像也都能提升另一方出现的概率。这里只研究成对的变量，以便搞清楚其中一个变量的出现是否会提升另一个变量出现的可能性。当遇到这种共同原因（当所有变量都被测量了之后）导致的混乱局面时，可以尝试用一个变量来解释其他变量之间的相关性，看看是否解释得通。这是哲学家们（包括 Suppes、Good 和 Reichenbach）提炼出的许多概率法的核心特征，也是使用计算法从数据中寻找原因的依据。

假设某种疾病会引起疲劳，并且人们通常会使用一种特定的药物来治疗这种疾病。我们可以设想一下，如果疲劳完全是这种疾病引起的，而且服用的药物并没有改善或加重疲劳的症状，那么服用这种药物所产生的变化是不会对疲劳症状产生影响的。如果病情保持稳定，那么我们将无法从其他变量上得到任何信息。这种将一个共同原因造成的各种结果分开来看的过程就叫“筛选法”。[22]

图 5-5a 中有药物和疲劳两个变量，而且似乎前者提升了后者出现的可能性。在服药的情况下，表示疲劳的灰色条形相对高一些；而在未服药的情况下，表示疲劳的灰色条形相对低一些，这意味着在服药的情况下出现疲劳症状的可能性比在未服药的情况下出现疲劳症状的可能性要大。然而，如果我们将有这种疾病的人和没有这种疾病的人分开来看（图 5-5b 和图 5-5c），就会发现无论他们有没有服药，这些人出现疲劳症状的概率都是一样的。一旦我们了解了疾病这个因素，就会知道这种药物不会改变疲劳症状出现的概率。



（a）未区分有无疾病　（b）有疾病　（c）无疾病

图 5-5　在不考虑参与者有无疾病的情况下（如图 5-5a 所示），服药和疲劳之间似乎存在相关性。然而，在考虑参与者有无疾病的情况下，服药和疲劳之间不存在相关性（无论是否服药，疲劳症状出现的概率都是一样的）

当一连串事件共同发生时，也可以将它们分开来看。我们假设上面这个例子的另一种情况：参与者得了某种疾病，医生给他开了一种药，而这种药确实有引起疲劳的副作用。如果这几个因素之间真正的关系是“疾病导致服药，服药导致疲劳”，那么我们就会发现这种疾病会提升出现疲劳症状的概率。为了能够采取更加直接的干预措施，我们通常都想找到最直接的因果关系。所以为了避免出现疲劳症状，病人应该停止服药或者改服其他药物。但如果我们错误地发现这种疾病和药物都会导致病人出现疲劳症状，就无法知道改服其他药物能否防止疲劳症状出现了。但如果我们再次将是否服药作为前提条件，疾病和疲劳之间的概率关系就不存在了。

没有什么方法是完美无缺的，筛选法能否成功还要取决于我们是否能找到真正的共同原因。如果经济衰退会导致通胀率下降和失业率上升，但我们却不知道经济是否处于衰退阶段，那就无法用筛选条件来证明通胀率和失业率之间的因果关系是虚假的。我们找到的关系到底是正确的还是错误的完全取决于是否找到了正确的变量集。第 6 章介绍计算法的时候还会遇到这个问题。有些方法确实可以明确有些情况是否存在隐藏的共同原因，但对于计算法而言，这仍然是一个尚未解决的问题。

问题还不止于此。有些情况并不存在可以分开两种结果的单一变量。假设 Alice 和 Bob 都喜欢上机器学习课程，而且都喜欢上下午的课。那么，无论我们是以上课的内容为条件还是以上课的时间为条件，都无法完全将 Alice 和 Bob 去上课的可能性区分开。如果我们只知道某个课程的上课时间，那么仍然能够从 Bob 是否上课的信息中推理出 Alice 有没有去上这堂课。因为 Bob 的上课与否可以向我们暗示这堂课的内容。在这个案例中，没有一个单一变量可以将 Alice 和 Bob 去上课的可能性分开。假如我们增加一个变量，这个变量只有在某个课程的上课时间是下午且课程内容为机器学习的情况下才能成立，那么问题就解决了。但是，我们首先需要对这个问题以及潜在的因果关系有所了解，这样才能知道是否需要增加这个复杂的变量——但我们并不总是能够做到这一点。到目前为止，我们还没有讨论过时间问题，因为我们已经理所当然地认为原因会出现在结果之前。但是单一变量能够解释一种相关性的情况不止一种，还有一种情况我们通常不会纳入分析，这种情况就是各种关系会随着时间而发生改变的情况。

要想知道筛选法的失败案例，可以回忆一下本节开头介绍的那几个不确定性案例。如果一件设备有问题，那我们可能无法完全将它造成的各种结果区分开来。举一个常见的例子：一个开关出了故障，它既能打开电视也能打开电灯，但并不是每次都能让电路闭合。如果电视打开了，灯也会打开，反之亦然。但是，并不是每次打开开关都能同时打开这两个设备。我们可以增加第四个表示电路是否闭合的变量来解决这个问题。但是，我们首先要对这个问题的结构有所了解（但我们并不总是能够做到这一点），这样才能知道是否有必要增加第四个变量。

要想解决这个问题，不是要去寻找某个准确的关系，而是要看看在其他因素保持不变的情况下，一个可能的原因对某个结果的影响究竟有多大。我们不要求在真正的原因保持不变的情况下，无论是否存在虚假因素，某种结果出现的概率都是完全一样的，而是要求某种结果出现的概率只会

发生很小的变化。“很小”这个词很模糊（这个值到底要多大才能算是因果关系呢），但是我们可以使用统计方法来衡量这些变化的显著性。

到目前为止，我们已经考察了所有能够让不是原因的事物也能提高某种结果出现的概率的方法。但是，有时一个真正的原因却可能无法提升某种结果出现的概率。某些原因会阻碍结果的出现，比如用来预防疾病的疫苗。这些问题很容易处理。我们可以根据某些原因会降低某种结果出现的可能性这一特征来重新定义原因，也可以将某种结果的反例作为我们关注的结果（即不出现疾病）。但在另一些情况下，正相关的原因似乎降低了某种结果出现的概率，或者对某种结果没有任何影响，这又是怎么回事呢？之所以会出现这种情况，主要原因在于采集数据的样本和我们针对变量使用的粒度级别。

5.3　辛普森悖论

假设你是一个病人，正要从两名医生中为自己挑选一名医生。对于某种疾病，A（Alice）医生治疗时的病人死亡率为 40%，而 B（Betty）医生治疗时的病人死亡率为 10%。如果只根据上述信息，那你可能会更倾向于选择 Betty 来为你治疗，但你其实并没有足够的信息来支持你的选择。

可能对于每一个具体的病人来说，虽然 Alice 的病人总体死亡率更高，但是她的治疗效果却更好。病人并不是随机被分给 Alice 和 Betty 的，有可能是其他医护人员推荐过来的，也可能是因为朋友、医评网站或广告上的推荐而来的。所以，如果 Alice 精湛的医术吸引来的是那些病情最严重并且最难治疗的病人，那么即便她是一名比 Betty 更好的医生，她的病人的总体死亡率也会很不乐观。

这件事的有趣之处在于，我们不仅能找到一个错误的因果关系，还

能找到与真正的关系正好相反的关系。比如我们发现 Alice 的治疗结果比较糟糕，但实际上她的治疗效果是更好的那个。如果我们考察的数据不是来自于随机的试验（病人被随机分配到不同的治疗小组中去），那么同样的情况也可能会出现在考察药品疗效的对比实验当中（我们将在第 7 章进一步讨论）。但是，这个实验在哪些病人服用哪一种药的问题上可能会存在偏差，要想解决这个问题，只能随机将病人分配到不同的治疗小组中去。比如说，如果每一个恶性肿瘤患者都采用 A 治疗方案，而那些更加容易治疗的病人都采用 B 治疗方案，那么由于接受 A 治疗方案的病人病情更加严重，所以 A 治疗方案的治疗效果肯定要更糟糕一些。选择偏差是导致我们难以从观察数据中进行推理的原因之一。我们可能会发现那些坚持锻炼的人比那些不爱运动的人更加长寿，但这也可能只是因为那些坚持锻炼的人比那些没有锻炼的或者不能锻炼的人更加健康而已。

因果关系可能会消失，或者表面上好像发生了逆转，这种奇怪的现象就是众所周知的辛普森悖论。[23]辛普森描述了这种情况中的数据具备的一些数学特征，并举了一个例子来进行说明：当我们分别考察男性参与者和女性参与者时，治疗方案是有效的；但当我们将所有参与者作为一个整体来考察时，治疗方案似乎没有任何效果。还有一些研究人员向我们演示了比这更加极端的情况：新的治疗方案对男性参与者和女性参与者的治疗效果都更好，但在整个参与者群体中却出现了更多的死亡病例[24]（如图 5-6 所示）。还有一些著名的案例，比如伯克利大学的研究生录取率（由于女学生申报的院系更具竞争性，所以录取率似乎要低一些）[25]和佛罗里达州的死刑率（被告的种族似乎是影响判决的一个因素，但实际上影响判决的因素是受害者的种族）。[26]

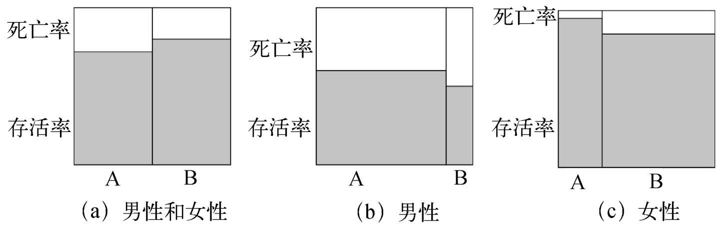


图 5-6　辛普森悖论演示图。图中 A 方案在男性和女性组中的效果都要更好一些，但 B 方案在整个群体组中的治疗效果似乎更好

在这些辛普森悖论的案例中，我们可以通过增加更多信息来解释这些虚假的关系，具体来说就是考察更小的群体。就那两个医生而言，一旦我们去考察一些健康状况相同或者风险级别相同的病人，就能发现其实 Alice 的治疗效果更好；就研究生录取率而言，可以分院系进行考察；就死刑率而言，可以按照受害者的种族来进行考察。这些实际上就是考察数据问题时使用的粒度级别。我们正在观察的概率实际上反映了概率背后的潜在关系，知道这一点有助于我们从数据中找到那些概率关系。在制定政策时，我们需要知道某个群体中出现的概率是否也适用于这个政策针对的群体。

问题的关键在于，我们要确定何时以及如何划分手中的数据，因为考察越来越细化的子数据集并不能够解决所有问题。整体数据集中不存在的、与直觉相悖的结果可能会出现在子数据集当中，而将这个数据进一步细分可能会导致这些结果再次逆转。在那个新药案例中，这种药在分别治疗男性参与者和女性参与者的时候效果更好，但在治疗整个参与者群体时，治疗效果似乎更糟糕。那么，我们也许应该相信这种药是有效的。这一点有些争议，因为辛普森自己说过这种新药“既然对男性参与者群体和女性参与者群体都有效，那我们就不应该认为它是对全人类毫无价值的药物”。[27] 但是，辛普森还举了一个这种解释不成立的例子。要想找到作

为区分条件的正确变量集，首先要对事件之间的因果结构有所了解，而如果我们研究变量集的目的就是掌握因果结构，那就麻烦了。[28]

这是很多因果关系问题的关键所在。我们绝不可能完全脱离因果关系的背景知识而去单独谈论因果关系问题，我们必须利用这种背景知识来选择数据分析的对象，并且用它来解释数据分析的结果。

5.4 反事实推理

如果你没有在我掷保龄球的时候弄出声响，我一定会击中的。如果外面再热一点的话，我一定会跑得慢一些。这两句话正在努力指出一个让事件这样发生而不是那样发生的突出因素。我们经常会从与事实相反的角度来谈论因果关系。当然，这一点我们也无法肯定——即使天气是完美的，我们也可能在跑步的时候岔了气，或者不得不停下来系鞋带。但我想说的是，在其他条件都不变的情况下，如果天气再好点的话，我可能会跑得更快一些。

这些话指出了一种必要性或一种造成差异的因素，而这种因素或必要性并不包含任何规律。通过休谟关于事件发生序列的规律性理论，我们仅能知道有些事件经常一起出现。而我们现在要试图阐明的是，在某种意义上，要想让这些事件按照它们已经发生过的方式再次出现，是离不开导致这些事件发生的那个原因的；如果这个原因没有出现，那么这些事件发生的方式就会大不相同了。这就叫反事实推理。大致来说，反事实推理就是一个这种形式的推断：如果 A 成立的话，C 也会成立。例如：如果我涂了防晒霜的话，那我就不会被晒伤了。

有趣的是，休谟的理论可以让人们同时运用规律法和反事实推理法来研究因果关系。休谟曾这样写道：“第一个事物出现在第二个事物之前，而且所有与第一个事物类似的事物出现之后，都会出现与第二个事

物类似的事物（规律性的定义）。”接着他又写道：“换句话说，如果第一个事物不出现的话，那么第二个事物也永远都不会出现（反事实推理的定义）。”[29]从休谟的文章中可以看出，他似乎认为这两句话说的是同一个意思，但实际上由此诞生的却是两种不同的因果研究方法。

反事实推理法（在休谟的启发下，由 David Lewis 正式提出）的基础是，要想让 C 能够引起 E，那么有两个条件必须成立：如果 C 没有出现的话，那么 E 也不会出现；如果 C 出现了的话，那么 E 也会出现。比如：如果我涂了防晒霜的话，那我就不会被晒伤了，而如果我没有涂防晒霜的话，那我就会被晒伤。这两个条件既包含了必要性又包含了充分性。当然，也有概率性反事实推理法，但是我们在这里不进行介绍。[30]

我们再回头看看编程马拉松的例子。也许这些程序员每次喝了很多咖啡之后第二天都会非常累。也许他们只有熬夜的时候才会喝很多咖啡。不管怎样，单凭这些事件出现的规律性，我们就能发现喝咖啡是导致疲劳的一个原因。但是，如果这些程序员不喝咖啡，他们第二天仍然会很累（由于熬夜所致，假设他们在没有咖啡因刺激的情况下仍然能够熬夜）。如果使用反事实推理法分析这个案例，那么喝咖啡就不再是引起疲劳的一个原因了。从理论上来说，这个方法让我们能够将那些可能偶然一起发生的因素和那些真正导致某种结果的原因区分开来。

现在你可能会想，我们如何才能真正了解将会发生的事情呢？这就是法律推理（我们将在后面探讨这个话题）的核心难题之一：我们能否肯定地说，如果你前面的车没有突然转向的话，你就不会急刹车，也就不会被后面的车撞了呢？也有可能你后面的驾驶员注意力不集中或者身体上有缺陷，无论如何都会撞上你的车呢？

与事实相反的事情通常是指单一的事件而不是普遍的特性（这些将在第 8 章详细讨论）。要想让这些单一的事件成为我们可以正式考察的事物，可以将这些事件与模型联系在一起。如果我们能够用一系列等式来表示一

个事件系统，就可以直接验证如果我们研究的原因不成立的话，那我们所关心的结果是否依然会出现。比如说，如果某种毒药是绝对致命的，那么只要吃了这个毒药就一定会死。当然，很多原因都能导致死亡，所以我们可以给其他原因设定一个值（真或假）。然后，我们就能看到如果我们改变毒药的值（真或假）会导致什么样的结果。如果我们将毒药的值设定为假，那么其他变量是否足以让死亡的值保持为真呢？这就是结构方程模型背后的基本理念。在这个模型中，每一个变量都是系统中其他变量的某个子集中的一个函数。[31]

然而，这种反事实推理法也不是完美无缺的。想想拉斯普金案吧！传说，他曾在吃蛋糕时喝下了有剧毒的葡萄酒（酒里所含的氰化物的量足以毒死五个壮汉），但并没有被毒死。结果有人朝他背后开了一枪，而他再次活了下来，接着又中了枪。最后，他被绑起来扔进了冰河之中，但他又自己解开了绳索！不过，最后拉斯普金还是被淹死了。那么，他的死因是什么呢？我们能否肯定地说如果他没有被下毒也依然会死呢？有可能是这个毒药过了段时间才发作呢？或者可能是这个毒药让他没有力气从河里游上岸呢？同理，中枪也可能会起同样的作用（以其他方式促成了他的死亡）。

这个例子里出现了好几个原因，而且其中任何一个原因都可能导致某种结果，像这样的例子很难进行反事实推理。这些案例都是超定的实例。超定就是多余因果关系的对称形式，比如在执行枪决的过程中，犯人被行刑队的多名成员射中；一个病人同时服用的两种药物都能引起某些副作用。在这两个例子中，如果其中一个原因没有发生（某一个队员没有开枪，或者病人没有服用某一种药），结果依然会发生。从反事实推理的角度来看，这个结果并不依赖于每一个原因。我们也可以放松这个条件说也许结果依然会出现，不过可能会有一点不同而已。也许副作用出现的时间可能会晚一些，或者没有那么严重。[32]

在超定的案例中存在一个问题，就是我们没有找到任何原因。但从概念上来讲，我们本来也无法真正确定某个具体的原因，而且每个原因似乎都在某种意义上合理地导致了某个结果的出现。我们再来看看这种情况：有两个原因同时存在，但是只有一个原因随时都能起作用，另一个原因只是一个备胎，只有在第一个原因不起作用的情况下才会起作用。假设行刑队的每一个成员只有在前一个队员开枪后没有杀死囚犯的情况下才会开枪。生物学中经常会出现这种类型的备胎机制，比如有两个基因能够产生同样的显性特征，但是其中任何一个基因都能抑制另一个基因的作用，即基因 A 抑制基因 B，以便只有在基因 A 不起作用的情况下基因 B 才会起作用。所以，显性特征 p 并不依赖于基因 A，因为如果基因 A 不起作用，基因 B 就会起作用。这个案例比前一个案例更麻烦，因为我们虽然可以通过直觉挑选出一个导致某种结果的因素，但是无法通过反事实推理法找到这个因素。在这个案例中，有两个或更多可能的原因都能导致某种结果，但实际出现的只有一个原因，这种类型的问题就叫优先权问题。

人们经常会区分“早到的优先权”和“迟到的优先权”。在早到的优先权案例中，只有一个原因会出现并完成整个过程，而另一个原因（在第一个原因没有起作用的情况下会起作用的第二个原因）则处于被抑制的状态。基因备胎案例正是这种情况。在迟到的优先权案例中，两个原因都会出现，但是导致某种结果的只有一个原因。行刑队案例正是这种情况：有一颗子弹会在其他子弹之前击中犯人，并在其他子弹击中犯人之前让犯人毙命。

在反事实推理方面，因果关系的具体结构还存在一些其他问题，尤其对于那些从反事实依赖链的角度来思考因果关系的人来说更是如此。如果存在一个反事实因果关系依赖链的话，那么据说这个依赖链中的第一个组成部分就是引起最后一个组成部分的原因。

比如说，《老爸老妈罗曼史》中有一集讲的是两个人在争执到底是谁

害他们错过了航班。Robin 觉得是 Barney 的错，因为 Barney 在去见他的路上在地铁站翻了一个旋转栅门，所以导致 Ted 被开了罚单，并且不得不在飞机起飞的那天早上去法庭受审。后来 Ted 分析了一个复杂的事件链（其中包括 Robin 导致 Marshall 脚趾受伤的事），然后觉得这是 Robin 的错，因为 Barney 之所以要跑马拉松（因此他在地铁站才需要帮助）完全是由 Robin 导致的。而 Robin 则觉得错在 Lily，因为她之所以会出现在 Lily 家（去睡一觉）并导致 Marshall 受到惊吓然后伤了脚趾，是因为她要排队购买特价婚纱。最后，故事的高潮是 Ted 认为这件事归根结底是他的责任。因为他发现了一枚罕见的幸运便士，然后他和 Robin 把这枚硬币卖掉了，用卖来的钱去婚纱店对面买了热狗。在这一集中，每一件事都有一个与事实相反的假设：如果 Ted 没有去法庭的话，他就不会错过航班；如果 Marshall 去跑马拉松的话，Barney 就不会需要 Robin 的帮助；如果 Robin 没有去婚纱店的话，Marshall 的脚趾就不会断；而如果 Ted 没有捡到硬币的话，他们就不会知道婚纱店在促销。[33]

在这种案例中，不同因果理论对真正的原因有着不同的观点。有些理论寻找的是引发这一系列事件并导致某种结果的最早的因素，还有些理论则想要找到最直接的原因。这些理论存在的问题是，我们可能会不断找到距离实际结果越来越远的事件。然而，还有更加麻烦的情况：某个事件通常会阻止某种结果的出现，但又会让这种结果以另一种方式出现，从而产生一个表面上的依赖链。比如，有一个见义勇为的人在一列火车前面救了一个摔倒在铁轨上的人，但这个人后来却在跳伞时摔死了。要是他在铁轨上没有被救的话他就不会去跳伞，从反事实推理的角度来看，他的死亡也就不会由跳伞来决定，因为跳伞是由他被救这一事件来决定的。这样一来，这位见义勇为的人似乎反倒成了导致他死亡的因素。在第 8 章，我们将研究人们在法律案件中是如何处理这类问题的。毕竟，如果一个被救的人后来酒驾并撞死了人，我们肯定不想让那个救人的人来承担责任。即便

是这个救人者的行为让这场交通事故成为可能，也不应该让他承担责任。尽管其中可能存在因果关系，但这还不足以让救人者承担法律责任。要想让救人者承担法律责任，还需要一个此案例中并不存在的条件：后果的可预见性。

5.5　观察法的局限性

我们再回头想想本章开头提到的那个统计数据，它声称有些因素能让人们有 98%的可能不会陷入贫困境地。现在你应该知道为什么从这个数据中去推理因果关系会那么难了。当只有观察数据时，我们永远都无法确定是否存在隐藏的共同原因，从而导致了一些表面上的因果关系。比如，我们可能会发现青少年时期玩暴力的电子游戏和成年后成为一个暴力的人之间存在相关性，但是青少年时期玩暴力的电子游戏可能完全是环境和基因因素导致的。同样，当我们只能观察而不能干预时，就必须将选择偏差考虑进来。比如说，我们假设参加锻炼的人对疼痛的忍耐力比不锻炼的人高。这并不能告诉我们锻炼是否真的能够提高人们对疼痛的忍耐力，也不能告诉我们这些坚持锻炼的人是否都是忍耐力高的人，因为他们更能适应不适感。但是，观察却能为我们以后的实验研究或探索因果机制（原因是如何导致某个结果的）背景知识的活动提供一个切入点。

注释

1. 在原文中，这个数据实际上是这样说的：“那些完成了高中学业、拥有全职工作并且在婚后才生孩子的人几乎无一例外都成了中产阶级。在这群人中，只有 2%的人成了贫困人口。”（Haskins 和 Sawhill，2009，9）。
2. 有证据表明，如果人们遇到的主要障碍是没钱，那么直接给人们现金可能会成为一个有效的干预措施。关于有条件现金转移支付和无条件现金转移

支付项目的效果对比，参见 Baird 等（2013）；关于无条件现金转移支付项目的回顾，参见 Haushofer 和 Shapiro（2013）。
3. 这是一项正在进行的研究。该研究追踪记录了弗雷明汉好几代居民的健康状况。
4. Mill（1843）。
5. 在计算方法领域，充分性还有另一个意思，指数据中包括哪些变量。
6. 由于人们假设的运行机制（即原因导致结果的方式）不同，可能会出现具有决定性特征的关系。
7. Corrao 等（2000）。
8. Nieman（1994）。
9. Mostofsky 等（2012）。
10. Snow（1855）。
11. Snow（1854）。
12. Snow（1854）。
13. Rothman（1976）。
14. Mackie（1974）。
15. Carey（2013）。
16. Dwyer（2013）。
17. Carey（2012）。
18. 想要了解关于统计功效的基本信息，参见（Vickers，2010）。
19. 不同国家的比例有细微的变化，但已经有很多大规模的研究使用了 SAH 登记数据。这些研究也给出了类似的数字（Korja 等，2013；de Rooij 等，2007；Sandvei 等，2011）。
20. Cherry 和 Oldford（2003）提出了用来表示概率的 Eikosogram 图。
21. Maurage 等（2013）。
22. 想要了解更多关于筛选的信息，参见 Reichenbach（1956）。
23. 在这些小组相互作用的基础之上，可能会出现似乎矛盾的结论。人们一般认为是辛普森（1951）首次普及了这个似乎矛盾的结论。然而，在辛普森之前，Yule （1903）也曾描述过这一现象。所以，有时这一悖论也称为 Yule-Simpson 悖论。这一悖论的发现也可归功于 Pearson 等（1899）。他们曾和 Yule 一起工作过。
24. Baker 和 Kramer（2001）。
25. Bickel 等（1975）。
26. Radelet 和 Pierce（1991）。

27. Simpson（1951），241。
28. 想要了解更多关于辛普森悖论的争论以及各种试图解决辛普森悖论的方法，参见 Hernan 等（2011）；Pearl（2014）。
29. Hume（1739），172。
30. 想要了解更多这方面的信息，参见 Lewis（1986b）。
31. 想要了解更多关于结构方程和反事实推理的内容，参见 Pearl（2000）；Woodward（2005）。
32. Lewis（2000）后来修改了他的反事实推理理论，以便将结果的出现方式考虑进来，以及在结果出现的事实不变的情况下，结果出现的方式可能也会不一样。
33. Rhonheimer 和 Fryman（2007）。

第 6 章　计算法

如何自动实现寻找原因的过程？

哪些药一起服用会产生不良反应？

针对这个问题，用随机试验来测试药品并不能给我们提供多少信息，因为这些试验往往会避免让参与者同时服用多种药物。虽然我们可以用模拟实验来预测药物之间的相互作用，但是这样的实验需要有大量的背景知识才能完成。我们也可以用实验的方法对一些药物组合之间的相互作用进行测试，但考虑到这种实验需要的成本和时间，它可能只适用于少数几种可能的药物组合。更糟糕的是，在数百万可能的药物组合中，只有少数几个组合的药物之间可能会出现严重的相互反应，而且这种反应可能只会在某些人群中出现。

一种药物上市之后，一些疑似不良反应的事件会被病人、制药公司和医疗服务机构报告给食品及药品管理局（FDA），并被输入数据库。[1] 所以，如果你服用了一种抗过敏药物，几天后心脏病发作了，那么你或者你的临床医师就可以把这一情况报告给 FDA。当然，这些报告里所说的情况通常都是未经证实的。可能某个人的心脏病发作实际上是与药物无关的血块引起的，但由于最近有新闻报道说出现了很多起该药物引发心脏病的事件，因此将这个人的心脏病发作解释为该药物引发的不良反应似乎就很合理了。很多情况都可能会导致数据出现虚假的因果关系。例如，病

人身上可能还有其他疾病引发了心脏病（比如未诊断出的糖尿病），这个数据本身也可能会出问题（比如样本被污染了或者症状被误诊了），而且事件发生的顺序可能并不是这样的（比如实验检测发现血糖升高了，但是血糖升高是在服药之前发生的）。很多真正的不良反应可能并未报告给 FDA，因为人们可能认为这些不良反应并不是服药引起的，也可能是因为病人在出现不良反应之后并没有去看医生，而且自己也没有将这个不良反应报告给 FDA。

即便有些报告所说的情况是错误的，它们仍然可以帮助我们形成新的有待检验的假设。如果我们想要通过实验来验证这些不良反应，比如让一组病人服用各种药物组合，或者让每个病人分别服用每种药物，那我们可能要耽误很长时间才能找到这些药物之间的相互作用，从而导致更多病人可能出现药物不良反应。相反，如果使用另一组来自医院的观察数据，我们就能准确地知道病人服用某种药物组合后会出现什么情况。斯坦福大学的一个研究团队正是这样做的。[2]他们使用的数据来自于 FDA 的不良反应数据库，发现同时服用某个降低胆固醇的药物和抗抑郁剂（分别是普伐他汀和帕罗西汀）可能会导致血糖升高。然后，他们又通过医院的记录比较了分别服用这两种药物和同时服用这两种药物的病人的实验室检测结果，发现病人在同时服用这两种药物之后，血糖升高的值比其他病人要大得多。

当然，我们无法确定病人有没有服用医院给他们开的药，也无法确定同时服用两种药物的病人和其他病人有没有什么不同。尽管这种类型的数据存在很多局限性，但研究人员使用了三家医院的数据进行验证，得出的结果都是一样的，还通过小白鼠的实验进一步验证了这一结果。[3]

在这项研究中，研究人员发现了两个相互作用的药物，但他们一开始并未假设这两种药物可能会相互作用，而是从数据中发现了这个假设。与之相反的是，我们目前所考察过的所有研究都是从某个具体的因果假设

出发，然后再对这个假设进行评估，比如糖吃多了是否会导致糖尿病这类研究。

但是，倘若我们并不清楚导致各种关系成立的因素是什么，比如医院的再入院率为什么会上升，或者是什么因素让人们访问各个网站的，那么我们要从网站上的交换信息、医院病历和网络搜索等数据集中了解什么样的信息呢？要何时了解呢？通过将计算能力和从数据中有效发现原因的各种方法进行结合，我们对数据的分析已经不再是一次只考察一个因果关系了，而是通过对数据的挖掘同时揭示多种因果关系。通过这些自动化的方法，我们还可以发现很多人们无法直接观察到的更加复杂的关系。比如，我们可能会发现一个让病人在中风后恢复意识的、由多个步骤（每个步骤又包含多个必要组成部分）组成的事件序列。

本章将考察从数据中寻找原因的方法。首先要讨论的是：什么样的数据适合用来推理因果关系？并不是每一个数据集都能让我们推理出正确的因果关系，所以我们将讨论必须在因果推理中做出什么样的假设（以保证推理出的因果关系是正确的），以及在这些假设不成立的时候，我们可以得出什么样的结论。虽然推理因果关系的方法有很多，但我们考察的主要是这两种类型的方法：试图找到一个模型来对数据进行解释的方法（本质上就是同时了解数据中所有的因果关系），以及重点对每一个关系的强度分别进行评估的方法。最重要的是，我们要认识到没有一种方法能永远胜过其他方法。尽管我们在计算方法上已经取得了巨大的进步，但这仍然是一个还在研究中的领域，而且在没有任何背景知识的情况下，我们还无法做到完美而又准确地推理出各种情况下的因果关系。

6.1 假设

在考察推理方法之前，我们还要讨论一下使用这些方法需要输入的

内容。这里所说的因果推理一般是指先选择一组被测变量（比如随时间变化的股票价格），然后使用一个计算程序来找出是哪个变量引起了哪个变量（比如 A 股票价格的上升引起了 B 股票价格的上升）。这可能意味着我们要找出每组股票之间的关系的强度，或者要找到一个模型来解释它们是如何相互影响的。这里所说的数据可能是指随着时间变化的事件序列，比如一只股票价格每天发生的变化，也可能是指某个时间点上的事件序列。在第二种情况中，我们考察的不是随着时间而产生的变化，而是各个样本之间的变化。比如在某个时间点上对一群人进行调查，而不是针对某些个体进行长期跟踪调查。

不同的研究方法假设出的数据也略有不同，但有些特征几乎对所有研究方法都是一样的，而且这些特征还会影响我们从数据中得出的结论。

6.1.1　无隐藏的共同原因

一个最重要且最普遍的假设可能就是我们已经测量了正在进行因果推理的变量中的所有共同原因。这在图示模型法（即将介绍）中也被称为因果关系的充分性。如果想要从一组变量中找出原因，那么我们必须确保测量了这些变量中的所有共同原因。如果咖啡因是真正导致睡眠不足与心率上升的原因（而且这也是睡眠和心率之间的唯一联系），那么如果我们不测量咖啡因的摄入量，可能就会得出错误的结论，在咖啡因导致的两个结果（睡眠不足和心率上升）之间找到联系。数据中缺少的原因叫作潜在变量。两个或两个以上的变量之间未测量到的原因可能会导致人们做出错误的推理，这样的原因被称为隐藏的共同原因或潜在的混杂因子，而由此导致的问题被称为混杂（在计算机科学和哲学文献中更为常见）和遗漏变量偏差（在统计学和经济学中更为常见）。这是观察性研究的主要局限性之一，也是大多数计算法输入内容的主要局限性之一。它不仅会导致人们在变量之间发现错误的联系，还会导致人们高估原因的强度。

现在对上面这个例子稍加改动，让咖啡因不仅能直接抑制睡眠，还能通过心率上升来抑制睡眠（如图 6-1 所示）。尽管心率上升会引起睡眠减少，但如果没有测量咖啡因的摄入量，我们可能会发现咖啡因的显著性比我们本应发现的显著性要高一些或低一些。也就是说，因为咖啡因会导致心率上升，所以心率高可以向我们透露咖啡因的状态（存在或者不存在）。我们将在第 7 章考察实验法是如何通过随机化来解决这个问题的。几乎每一个使用观察数据的方法都必须假设不存在隐藏的共同原因，但实际上我们只有在极少数的情况下才能够保证确实不存在隐藏的共同原因。

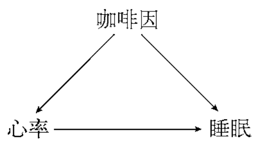


图 6-1　咖啡因是心率上升和失眠的共同原因，但是心率也会直接影响睡眠

注意，我们并不一定非要假设每一个原因都要测量到——我们只需测量那些共同的原因。如图 6-2a 所示，图中咖啡因不仅引起了睡眠的变化，还引起了心率的变化，而白酒也同样引起了睡眠的变化。如果没有关于白酒摄入量的数据，那我们将无法找到引起睡眠变化的原因，但也不会因此就在其他变量之间找到错误的关系。同样，如果咖啡对睡眠的影响是通过一个中间变量引起的，它们之间的关系差不多是咖啡因引起心率上升，而心率上升又导致睡眠减少（如图 6-2b 所示），那我们如果不测量心率，最多只会找到一个间接的原因，而不是一个错误的因果关系结构。因此，并不一定非要观察到因果关系链中的每一个环节。

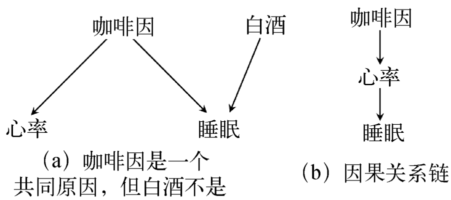


图 6-2　即使没有测量白酒（左边）和心率（右边），也不会混杂咖啡因和睡眠之间的关系

有些计算法试图寻找什么时候可能会存在缺失的原因，或者试图在某些情况下寻找这个原因本身，以此来避免要测量到所有共同原因的假设。然而，这一点通常只有在相当严格的条件下才能做到，而且在复杂的时间序列数据中难度会更高。[4]

如果我们并不知道所有的共同原因都被测量到了，也不能使用这些方法来找到这些共同原因，又该怎么办呢？本章将要考察的图示模型法中有一个办法，就是找到与这个数据一致的所有可能的模型，包括那些带有隐藏变量的模型。比如说，如果我们在睡眠和心率之间发现了一个表面上的因果关系，并且知道这两个变量之间可能存在某些未测量到的共同原因，那么一个可能的模型就会包含一个（能够引起这两个观察到的变量的）隐藏变量。这种方法的好处在于，所有能够解释这些数据的模型之间可能会存在一些共同的联系。这样一来，即便存在多种可能的因果结构，我们依然能够找出一些可能的联系。

人们对因果推理结果的信心与他们对所有潜在原因的测量程度成正比，无一例外。在将来验证这一结论的实验研究工作中，我们根据观察数据推理出的结论可以作为这些研究工作的出发点。

6.1.2　典型分布

除了要确保找到了正确的变量集，我们还需要确保观察到的内容反映了观察对象的真实行为。从本质上来说，如果没有报警系统就会导致抢劫

案的发生，那么我们的数据需要确保抢劫案的发生完全依赖于是否安装了报警系统。我们已经考察了几个数据不具代表性的案例：考察有限范围内的数据导致我们发现学习和 SAT 成绩之间没有任何相关性（见第 3 章）；辛普森悖论表明，根据考察数据的不同（整体数据或分别考察男性参与者和女性参与者的数据），药物和药效之间的因果关系会消失或发生逆转（见第 5 章）。

我们还考察过一个案例，这个案例向我们展示了各种关系是如何相互抵消从而导致了一个没有相关性的因果关系的。在第 3 章中，跑步和体重下降之间存在两种关系，一种是跑步对体重下降有积极影响，另一种是跑步对体重下降有消极影响，因为跑步同时也会导致食欲的增加。如果搜集到的数据分布得不好，我们可能就会发现跑步和体重下降之间没有任何关系。因果推理取决于真实的依赖性关系，所以我们通常要假设这种类型的相互抵消是不会发生的。这种假设通常被称为忠实性原则，因为那些不能反映真正的潜在因果结构的数据在某种意义上是“不忠实的”。

有些人认为这种违背忠实性原则的现象并不常见，[5] 但有些系统（比如生物系统）的结构方式就几乎确保了这种现象一定会发生。当很多基因都能产生同一种显性特征时，即便我们让其中一个基因不起作用，这个显性特征依然会出现，这就导致人们认为原因和结果之间似乎并不存在依赖性。很多需要保持平衡的系统都有这种类型的备用原因。

有时候甚至不需要有真正的抵消效应或无依赖性的因果关系就能违反忠实性假设。因为在实际研究中，大部分计算法都要求我们设定一个统计临界值，用于界定变量之间的关联是否可以被接受（使用 P 值或其他标准）。所以，某个结果出现的概率无须与其在某种原因下出现的概率完全相等，只要两个数值的差别足够小，能够保证结果仍处于可接受的范围之内就可以了。比如说，跑步之后体重下降的概率可能与不跑步也体重下降的概率并不相同，但如果二者的差别极小，可能就会导致跑步与体重下降

之间的关系违反忠实性假设。[6]

选择偏差也会导致分布的数据无法反映各种真实关系。比如这里有一份来自医院的数据，其中包括各种诊断记录和检查结果。然而，有一项检查十分昂贵，所以只有在病人出现十分罕见的症状并且无法通过其他方式确诊时，医生才会让病人去做这项检查。结果，绝大部分检查结果都是阳性的，但我们无法从这些观察数据中得知这种检查结果呈阳性的真正概率。因为只有在检查结果呈阳性的可能性很大时，医生才会让病人去做这个检查。我们观察到的通常是医学检测中一个非常有限的范围，比如说有些测试可能只针对那些病情最严重的病人（比如在重症监护室的有创血压监测）。观察到的数值范围也只包括那些病情严重到需要用这种监测的病人的数据。这意味着如果我们在这个有限的群体中发现了一种因果关系，也不代表这种因果关系在整个人群中也能成立，这就是我们面临的问题。同样，由于这种样本缺乏变化，我们可能无法找到事物之间的真正关联。

这与数据缺失的问题有关。变量的缺失可能会导致关系的错乱，而测量数据的缺失能够产生一种无法反映真正潜在关系的分布形式，从而同样导致人们做出错误的推理。数据的缺失通常都不是随机删除数据集中的数据导致的，而是取决于其他变量有没有被测量到。比如说，在住院病人的医疗程序中，可能需要断开一些监测器（导致数据记录中出现空白），或者某个设备故障可能会导致有些数据没有被记录下来。当血糖超出正常范围时，人们可能会更加频繁地测量血糖值，所以测量数据中的大幅空白和实际测量值之间并非是毫无联系的，而测量出来的数据也可能会偏向极端区间。由于某个隐藏的原因所导致的数据缺失可能会引发混乱，而设备故障则可能意味着其他相邻的测量结果也有问题（并可能导致结果出现偏差）。

实际上，我们只能假设在样本足够大的时候，样本的分布会反映数据背后的真实结构。如果我给一个朋友打电话，然后我的门铃立刻响了起来，我无法断言这种现象是否还会再发生。但如果我注意到这种现象出现了 5 次或者 15 次呢？假设，通常随着数据集大小的增长，我们会越来越接近事件的真实分布情况。如果你只抛了几次一枚公正的硬币，那你可能不会看到正面朝上的次数与反面朝上的次数相一致，但当你抛硬币的次数接近无穷次时，正面朝上和反面朝上的次数比会接近 1∶1（50/50）。这里所说的更多的数据指的是见到一连串罕见事件的机会减少了。这一连串的罕见事件并不能反映事件背后真正的概率，比如掷骰子时连续三个六这种事件。

我们在进行因果推理时也会做出同样的假设：假设我们有足够多的数据，假设我们看到的是由某个原因引起的某个结果出现的真正概率，而不是一个异常现象。需要注意的是，对于有些系统（比如那些非稳定性系统）而言，即便是一个无穷大的数据集也无法满足这个假设的要求，所以一般情况下，我们必须假设这些关系是不会随着时间的变化而改变的。前面说过，非稳定性指的是那些像股票平均日收益一样随着时间而改变的特性。在图 6-3 中，打折销售（虚线时间序列）和热巧克力销量（实线时间序列）在整个虚构的时间序列中几乎没有任何相关性，但是它们在阴影期（代表冬季）却是高度相关的。所以，如果我们使用了所有的数据，就不会发现打折销售导致了热巧克力的销量上升。而如果我们只使用冬季的数据，则可能会发现这两者具有很高的相关性。值得注意的是，更多的数据并不能解决这个问题，我们需要使用其他方法来处理这个问题，具体内容请参见第 4 章。[7]

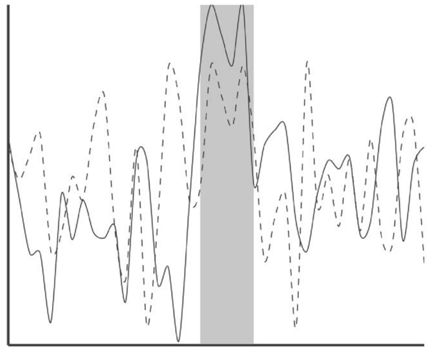


图 6-3　两个变量之间的关系随着时间的变化而发生改变，而且它们只有在阴影期才具有相关性

6.1.3　正确的变量

大部分推理方法都是为了找到变量之间的各种关系。如果你手上掌握的是金融市场的数据，那你研究的变量可能会是各个股票；如果在政治学领域，那你研究的变量可能是竞选捐款额和通话量。一般情况下，我们要么从一组已测量的事物出发，要么出去做一些实地测量活动，而且通常会将我们测量的每一个事物都看作一个变量。

我们不仅需要测量正确的事物，还需要确保描述这些事物的方式是正确的——这一点一定要明确。在组织信息的过程中，除了要处理是否保留某些信息这种简单的问题，还需要做出很多选择。在某些研究中，肥胖和肥胖症可能属于一个类别（所以我们只要记录每个个体是否肥胖或患有肥胖症就可以了），但是对于那些致力于治疗肥胖症患者的研究来说，对肥胖和肥胖症的区分可能就至关重要了。[8]

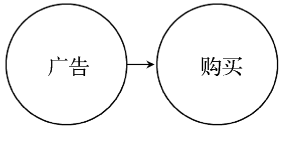
甚至只通过询问这样的分类，我们就已经做出另一个选择了。通过测量体重我们获得了一组数值，这些数值被映射到了不同的类别中。重要的

可能不是体重，而是体重的变化或变化的速度。我们可以计算每日的体重变化量或每周的体重变化趋势，无须使用最初的体重数据。由于结果总是相对变量组而言的，所以无论我们做出的决定是什么，它都会改变我们发现的结论。剔除一些变量可能会让其他原因看似更显著（比如剔除一个备用原因可能会让留下的那个原因的影响看似更强大），而增加一些原因则会降低另外一些原因的显著性（比如增加一个共同的原因能够剔除我们在各种结果之间误加上的关系）。

回想一下本章开头的那个例子。单独服用两种药物不会导致血糖升高，但同时服用这两种药物就对病人的血糖值产生了显著的影响。在各个变量和各种生理测量值（比如血糖值）之间进行因果推理可能无法找到任何关系，但如果同时考察这两个变量和生理测量值，我们就可以找到这种不良反应。在这个案例中，正确的变量就是同时服用这两种药物。要想找到正确的变量不是一件容易的事，这也是我们可能无法从一些数据集中做出重要推理的原因之一。

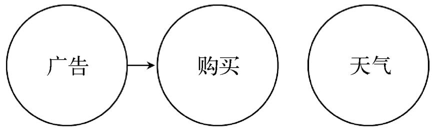
6.2 图解模型

为了向别人描述某个因果关系，或者为了理解各个事物是如何组成一个整体的，我们常常会画一张图。这些图形实际上可以和哲学家们的因果概率理论联系在一起。下面这个图形展示的是一个变量出现的概率是如何受另一个变量影响的。



这个图形首先告诉我们广告和购买行为之间存在某种关系。然后又

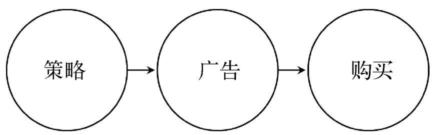
告诉我们这个关系是单向的，即广告影响购买行为，而不是购买行为影响广告。现在，我们再加上另一个变量。



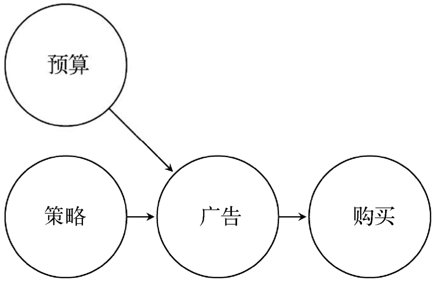
如果我们想要预测是否会发生购买行为，需要知道些什么呢？这些变量之间的连接方式告诉我们，我们唯一需要知道的就是，是否有人看到这个广告。从视觉上来说，天气位于图形的右侧，和前两个变量没有任何联系，而且它和购买行为之间也没有箭头，这就意味着我们不能使用天气来影响或预测购买行为。

只要知道一个变量的直接原因就能够预测这个变量，这个前提条件被称为因果关系中的马尔可夫条件。[9] 更严格地说，在变量的原因已经给定的情况下，变量是独立于它的非衍生物的（衍生物指的是由变量导致的结果，以及由这些结果导致的结果，等等）。[10] 这里的箭头是从原因指向结果的，所以直接原因就是那些通过一个箭头与某个结果联系在一起的原因。

为了说明直接原因的重要性，我们将做广告的原因也增加进来。

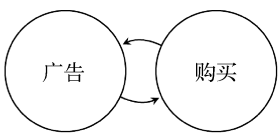


如果营销策略只能通过广告来影响购买行为，那么购买行为发生的概率则只取决于广告——导致购买行为的直接原因。一旦广告这个值确定了，那么它的产生方式就不重要了。即使我们发现了很多导致这个广告产生的原因，也不会改变我们预测购买行为所需的信息。这是因为所有的原因对购买行为的影响都要通过广告来实现。以下面这个图为例。



根据这幅图，如果我们想要知道关于购买行为的信息，就无须知道这些广告是否来自同一个策略，或者是否属于某一个预算庞大的广告攻势。要想知道购买行为是否会发生，我们只要知道广告有没有播出就可以了。这和第 5 章介绍的筛选法的思路是一样的。从理论上说，如果我们能够直接干预广告活动，那么无须对营销策略或预算做任何调整也能让购买行为发生变化，因为购买行为完全是由我们设定的广告值决定的。然而，我们几乎不可能真的只单独干预某一个变量而不改变图中的其他变量（第 7 章将详细介绍这一内容）。我们不可能像变戏法一样把广告变来又变走，而且这些干预措施还会导致各种预想不到的副作用。

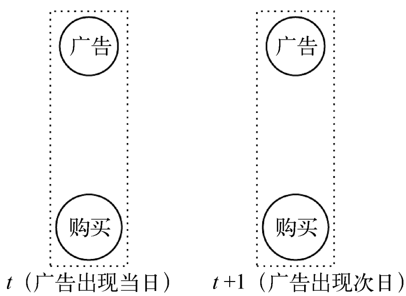
然而，这种类型的图形并不能反映每一种可能发生的情况。更多的购买行为也可能会导致广告投放量的增加或者导致营销策略发生变化，从而在图中形成一个循环。我们将要介绍一种叫作贝叶斯网络的图解模型，[11] 它是一种有方向的非循环图形。非循环指的是图形中没有循环，所以下面这个图形不包括在这种模型范围之内。



假设你沿着非循环图上的一条路径往前走，那你永远都不可能回到你出发的那个点。在用这些图形结构简化概率计算活动时，非循环图的这

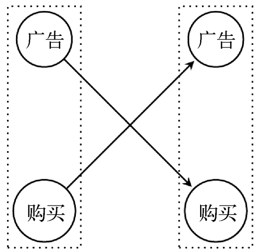
一特征至关重要。举个简单的例子：在每个变量非真即假的情况下，我们想知道购买行为和广告同时出现的概率。如果没有循环，当广告和购买行为中间只有一个箭头时，这两个事件同时发生的概率就是在出现广告的情况下发生购买行为的概率乘以出现广告的概率。[12] 因为购买行为是由广告决定的，所以我们只需知道购买行为在出现广告时发生的概率就可以了。然后，我们还要考虑广告真正发生的概率。比如说，人们在观看某个广告之后发生购买行为的概率为 1，但是出现广告的概率要低一些，比如 0.01，那么两者一起发生的概率就是 0.01。

但如果广告和购买行为之间存在一个反馈循环，那么广告出现的概率也会依赖于购买行为发生的概率。如果我们想要这个影响同时发生，就会加大概率计算活动的难度，但通过增加时间变量可以解决这个问题。我们可以假设在某个时间段发生的购买行为对广告产生的影响并不会立即出现，而是存在一定的延迟。为了表示这一现象，我们需要多个图形。



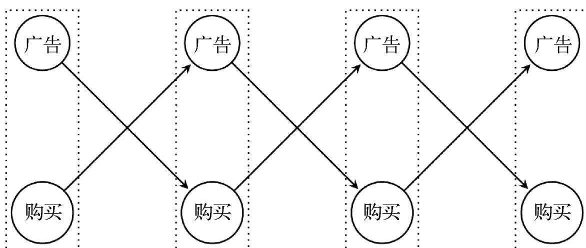
图左表示的是各个变量在 t 时刻的联系，图右表示的是各个变量在下一个时刻（即 $t+1$）的联系。在这两个图形中，广告和购买行为都没有连接在一起，因为它们不会立即对对方产生影响。就每个时间点而言，这两个图形中都是一个贝叶斯网络，所以它们是不可能有循环的。但只要我们不把这两个变量都放在同一图形中，广告和购买行为之间就可能会出现

即时效应，反之亦然。接下来，我们跨越时间将这些图形连接起来，以表示变量之间的反馈作用。



t（广告出现当日）　*t*+1（广告出现次日）

然后，这个图形结构就会随着时间的延续而不断重复，每一个时间点的购买行为都取决于前一个时间点的广告值，反之亦然。



t（广告出现当日）　*t*+1（广告出现次日，以此类推）　*t*+2　*t*+3

这种图形叫作动态贝叶斯网络，但图形结构本身并不会随着时间的变化而真正发生改变。[13]有的结构更加复杂，会出现多重时间间隔，并且变量之间的联系也不一定会立即在下一个时间点出现。有的结构可能会出现更长的时间间隔，比如接触某种病毒和出现感染症状之间的时间间隔。注意，随着变量及时间间隔数量的增加，我们推理这些结构的复杂程度也会大大提高。

6.2.1　图解模型在什么情况下会表示因果关系

尽管我们可以用图形来表示因果关系，但这并不意味着我们绘制的

或者知道的每一个图解模型表示的都是因果关系。到目前为止，我们只用图形表示了一个事物出现的概率是如何受另一个事物出现概率影响的。但是，我们也可以用图形来表示如何通过声音特征实现语音识别，如何根据内容来过滤垃圾信息，以及如何通过图像来识别人脸。此外，可能有多个图形都与一组概率关系一致（即多个图形都能用来表示同一组依赖性关系）。

我们怎样才能知道某个图示模型表示的是因果关系呢？这个问题的答案主要藏在那些将图形和（目前所讨论的）各种理论连在一起的假定之中。将图示模型用于因果推理的主要研究者不仅有哲学家，还有那些将因果关系哲学和图示模型结合在一起的计算科学家。

假设广告不仅能够引起购买行为，还能提高品牌认知度（如图 6-4a 所示）。如果我们没有用来表示广告的变量，还要试图从一组数据中推理出变量之间的关系，那我们可能会发现如图 6-4b 所示的图形，让我们错误地认为购买行为提高了品牌认知度。回想一下本章前面介绍的无隐藏的共同原因的假设或者原因充分性假设，我们在这里需要借用这些假设来避免出错。一般来说，任意数量的变量中都可能有一个共同的原因，如果这个原因没有被测量到的话，我们就无法保证由此推理出的关系是正确的。

如果广告变量表示的是“是否在电视台购买了广告空间”，但真正的原因却是消费者看到广告的次数，情况又会怎么样呢？与之前一样，我们需要找到正确的变量。因果关系可能会包括各种复杂的条件组合：也许抽一次烟不太可能导致肺癌，但是连抽很多年就很有可能导致肺癌；药物通常都有不同等级的毒性，所以服用 5 毫克药物可能不会产生不良反应，但是服用同样的药物 50 毫克却可以致命；西柚本身是无毒的，但它可以与很多药物相互作用，从而产生严重的不良反应。如果变量仅仅是抽烟（而不是烟龄）、是否服用某种药物（而不是服用该药物的剂量）和食用西柚（而不是在服用某种药物时食用西柚），我们可能无法找到这些因果关系，也可能会发现一些错误的关系。

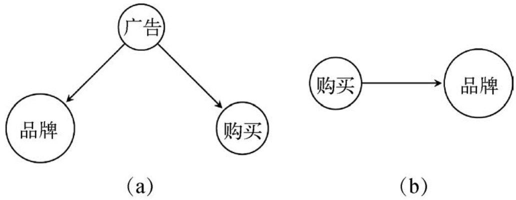
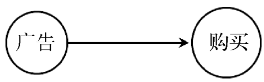


图 6-4　左图反映了事件之间真实的结构。如果没有观察到广告这个变量，我们可能就会发现右图这个错误的结构

这些结构表示的是概率关系，我们从中能够得知需要获得哪些变量才能预测出其他变量出现的概率。但若要想真正算出这个概率，我们还需要另一条信息。

一个贝叶斯网络包括两部分：结构（各个变量之间的连接方式）和条件性概率分布组合。简单来说，这些组合不过是一些表格。这些表格让我们能够在给定原因变量值（真或假）的情况下得知一个变量的两个值（真或假）出现的概率。针对广告和购买行为的那个图形，我们有一个两行两列的表格。

	购买行为为真	购买行为为假
广告值为真	0.8	0.2
广告值为假	0.3	0.7



每一行的概率和为 1，因为不管广告的值是什么，购买行为必须有一个值，而这些值出现的概率之和必须为 1。每一列的概率和不为 1，因为它反映的是购买行为的某一个值在广告值为真和广告值为假这两种前提下出现的概率。这个简单的图表还不完整，还需要另一个反映广告概率的表格。我们已经知道了在广告值给定的情况下如何确定购买行为发生的概率，但是怎样才能找到广告出现的概率呢？表示广告概率的表格中只有两个数字，因为广告在这个图形中是没有父级元素的，而且它出现的概率也

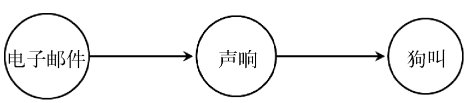
不取决于任何事物(就像抛硬币时出现正面朝上或反面朝上的概率通常也不取决于任何变量的值)。

对于贝叶斯网络中的每一个节点来说，我们都会有一个类似的表格。知道这个网络结构可以极大地简化我们的计算工作，因为每一个变量的值都是由其父级元素决定的。相反，倘若我们对变量之间的联系一无所知，就不得不将每一个变量都包含到表格的每一行之中。如果存在 N 个可以为真或为假的变量，那么我们就会有 2^N 行。我们既可以从数据中了解变量之间的结构和各个变量出现的概率，也可以根据我们了解的信息构建一个结构，以此来了解各个变量发生的概率。

无论在哪一种情况下，我们都要保证这些数据能够准确反映变量之间真正的依赖关系。这就又回到了典型分布假设或忠实性原则的问题上了。比如说，某些广告不可能既通过某种方式增加购买行为，又通过决策疲劳等因素减少购买行为。如果出现这种情况，我们可能会发现广告和购买行为之间没有任何依赖关系，虽然它们在真正的结构中之间是存在依赖关系的。另外，如果数据点太少的话，我们可能也无法准确地找到各个变量出现的正确概率。

有些情况下，忠实性原则可能也无法实现，比如第 5 章讨论过的辛普森悖论。其中一个案例告诉我们，由于我们对数据的划分方式不同(比如是研究所有病人，还是只研究男性病人或女性病人)，如果存在分组偏差(比如服用 A 药品的女性比服用 B 药品的女性多)并且结果也不一样的话(比如不管有没有服药，女性病人的表现都比男性病人好)，我们可能会看到并不存在的独立性。

另一种复杂的情况是出现决定性关系。比如说，每收到一封电子邮件，我的电脑都会发出声响，而电脑的声响又会让我的狗汪汪乱叫。



如果在出现声响的情况下，狗叫的概率为 1，而在出现电子邮件的情况下，电脑发出声响的概率也为 1（所以当这两个事件的原因出现时，这两个事件也一定会出现），那么声响是不会让电子邮件和狗叫成为独立性事件的，尽管这个结构告诉我们它们应该是相互独立的。假设你只知道是否收到了电子邮件。如果收到了电子邮件，电脑就会发出声响，而电脑发出声响后狗就会乱叫，所以你就能由此知道其他变量的状态。因此，你可能会错误地发现电子邮件直接导致了另外两个变量的出现。这个问题不仅是图示模型中存在的问题，也是大部分概率法中的一个难题。

总的来说，在下列假设中，图解模型表示的是因果关系。

- 一个变量的概率只取决于引起这个变量的原因（因果关系中的马尔可夫条件）。
- 所有共同的原因都要测量到（充分性原则）。
- 我们使用的数据准确地反映了变量之间真正的依赖关系（忠实性原则）。

还有一些隐含的假设（比如充分的数据、变量的描述必须正确等）也能保证因果推理的正确性，但是上述三个假设是最广为人知的，也是表示因果关系和不表示因果关系的图形之间最主要的差别。

6.2.2 从数据到图形

假设我们有一些关于某个公司雇员情况的数据。我们知道他们的工作时间、休假信息、部分生产指标等信息，怎样才能找到这些因素之间的因果关系网络呢？[14]

我们可以找一个指标来衡量一个模型对数据的描述能力，然后搜索可能的模型，找到在这个指标下得分最高的模型。这种方法叫作搜索评分法。[15]如果休假导致生产力提高是这个数据中的唯一关系，那么带有这样一个（从休假指向生产力的）箭头的模型应该比包含其他关系的模型或者

箭头方向相反（从生产力指向休假的箭头）的模型得分高，即图 6-5a 的得分应该比其他图形的得分高。因为只有三个变量，所以我们可以列出所有可能的图形，逐个测试，然后再做出选择。

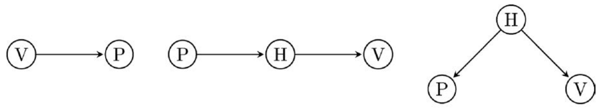


图 6-5　如果实际情况是 V（休假）导致 P（生产力），那么第一个图形的得分应该是最高的

要想从中做出选择，我们还需要用一个方法来计算哪个图形与数据更相符。用来评分的函数有很多，[16]但从根本上来说，除了要避免将图形和特定数据集中的噪声与特征进行匹配，我们对数据的描述程度也存在一个平衡点。我们可以通过一个非常复杂的结构来完美解释数据集中的每一个点，但我们想要找到一个模型来描述各个变量之间更为普遍的关系，而不是解释数据中的每一处噪声。

因此，当图形变得越来越复杂时，有些因素可能就很难解释了。然而，我们不能从所有可能的图形中进行选择。一个仅有 10 个变量的数据集就有 10^{18} 种可能的图形，[17]这些图形的数量是美元流通量的 100 万倍以上。[18]更不要说 S&P 500 指数中所有股票之间的各种关系了。只要 25 个变量，我们得到的所有可能图形的数量（超过 10^{110}）就会让宇宙中所有原子的数量（估计只有 10^{80}）相形见绌。[19]

没有任何方法能让我们一一测试这些图形，但其实也不需要一一测试它们。我们可以随机想出尽可能多的图形，然后再选出其中最好的一个。由于可能出现的图形数量太多，所以我们碰巧选中最好的那个的可能性不大。因此，我们需要为那些计算程序提供一些线索，告诉它们哪些图形更值得研究。

假设我们一一测试了图 6-6 中的前三个图形，然后发现图 6-6c 的得分最高。接下来最好的策略是去研究与这个图形相近的其他图形，而不是随机想出第四个图形。我们可以增加一个箭头、改变箭头的方向或者删除一个箭头，来看看图形的得分是如何变化的。也有可能最好的图形其实是图 6-6d 所示的图形，但由于我们使用了上述策略，一直在优化第三个图形，并且在找到真正的结构之前就已经停止了测试工作，所以我们永远也没有机会测试到第四个图形。如果我们不能测试到每一个图形，就无法确保最好的图形已经被测试了。图 6-7 向我们解释了这种局部最优化的问题。如果 y 轴代表图形的得分，而我们只测试标记点周围的图形，由于那个区域中最高的点就是这个标记点，那我们可能就会认为那个点就是所有图形中得分最高的点，这就叫作陷入局部最优化陷阱。虽然我们在某个区域得到了最高分，但这却不是所有可能的分数中最高的分数。

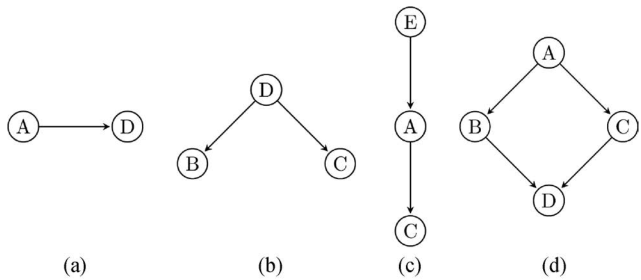


图 6-6　图中 A、B、C、D 为变量。图（a）（b）（c）展示的是可能被测试到的各种可能的图形，图（d）展示的是变量之间真正的结构

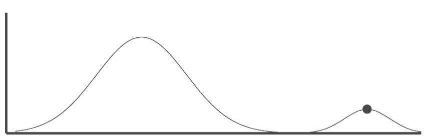


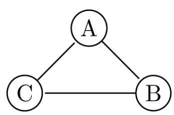
图 6-7　局部最优化示意图

为了解决这个问题，用于寻找因果结构的各种算法使用了更加巧妙

的方法，以此来限制需要测试的图形组合，并且尽可能多地去探索搜索空间的各个领域。如果我们知道性别只能是原因而绝不会是结果，那就可以避免测试所有将性别当成结果的图形。如果我们对要寻找的结构样式有一定的了解，那么就能为整个图形组合设计出一个概率分布图，并且可以用它来引导我们找到那些更有研究价值的各种可能出现的结构。[20]

除了去搜索那个数量多得可怕的潜在图形集以外，我们还可以使用变量之间的依赖性来建构那个得分最高的图形。约束法就是这样做的。它不断重复测试变量之间的独立性，并在测试结果中增加、减少图形中的箭头，或者改变图形中箭头的方向。其中有些方法是每次增加一个变量，还有一些方法一开始就已经将所有的变量连接在了一起，然后再一个一个地删除箭头。[21]

以下图为例。图中三个变量之间所有可能出现的联系都已经绘制出来了。



如果我们发现在给定 C 的情况下，A 和 B 是相互独立的，那么就可以删除它们之间的连线，然后继续寻找变量之间的其他关系，看看还能删除哪些连线。测试的顺序也很重要，前面出现的一个错误可能会导致后面出现更多的错误。在使用真实数据的案例中，我们看到变量之间完全相互独立的可能性不大，而我们需要判断的是，应该在什么时候接受或拒绝变量之间相互独立的假设。如果在给定 B 的情况下，A 出现的概率和 A 本身出现的概率完全一样，那么这两个变量之间就是相互独立的。但是，也有可能在给定 B 和 C 的情况下，A 出现的概率和在只给定 C 的情况下 A 出现的概率十分相近，但不完全一样。在实际研究中，我们需要选择一个统计学上的门槛（临界值），来决定是否接受基于这些测试而提出的条件

独立性结论。此外，由于我们需要做的测试数量庞大，所以很有可能会受到多重假设中很多问题的影响（参见死三文鱼实验）。[22]

6.3 衡量因果关系

有一种推理方法是去寻找一个与数据一致的或者能够对数据做出解释的模型。但是，这种方法在计算上可能会十分复杂，而且我们有时只想知道我们测量的所有变量中部分变量之间的关系。比如，我们可能只想知道生产力提高的原因，那就不需要一个包含所有被测变量的完整模型。随机试验解决的正是这种问题（比如某种药物对死亡率有什么影响），但它也有局限性（详见第 7 章），并不适用于所有的情况。

还有一种推理方法主要研究的是量化各种因果关系强度的问题。如果休假可以提高生产力，但生产力不能导致休假，那么休假作为提高生产力的原因的强度应该很高；反之，生产力提高作为放假的原因的强度应该很低。尽管相关性是对称的，但在衡量因果关系的显著性时，需要利用这些关系中的非对称性特征。在某种意义上，因果关系的显著性应该与原因对结果的解释程度相称，与原因作为一种干预手段能够带来某种结果的有效程度也相称。如果休假只能偶尔提高生产力，而加班总是能够提高生产力，那么作为生产力提高的原因，加班的强度要高于休假的强度。同样，如果强迫人们休假是提高生产力的有效策略，而强迫人们加班不是提高生产力的有效策略，那么休假就会再次成为提高生产力的一个更为显著的原因。

如果休假能够提高生产力只是因为休假可以让员工在这家公司待得久一些，且有经验员工的生产力更高，那么我们想要知道的是，经验对于提高生产力的重要性是否高于休假。也就是说，我们想要发现最直接的原因（这些原因在图形中是父级原因，而不是更遥远的祖父级原因）。

如果可以用一种方法（完全独立于引起其他任何变量的原因）去评估生产力提高的原因，那么我们可以做更少的测试，并且可以同时进行这些测试，从而大大提高计算这些事情的计算程序的速度。这种方法对我们有着很大的吸引力——这意味着我们无须再使用近似法（比如只研究一个子集而不是所有的图形），因为在近似法中，同一个程序运行几次，每次得出的结果可能都不一样；这还意味着这些计算将会变得十分简单，我们可以用精确法来进行计算。

这种方法的局限性在于，如果没有一个结构来展示所有变量之间的联系，我们可能无法直接使用这些结论来进行预测。假设我们发现党派的支持会让参议员们投票支持某些法案，而这些参议员所属选区选民的支持也会起到同样的效果。这并没有告诉我们这两种支持是如何相互作用的，也没有告诉我们如果这两种支持相加，是否会导致参议员支持某个法案的决心更强。要想解决这个问题，可以去寻找更为复杂的关系。我们并不是要使用所有测量过的变量，而是要去建立各种联系（政党和选民对提案的支持）、了解某个变量值必须为真的时长（锻炼一天、一个月、一年等）并考察一系列事件的先后顺序（先服用药物一或者先服用药物二）。我们在此就不详细论述了，但是，确实有些方法可以用来表示或者测试这类复杂的关系。[23]

6.3.1　概率与因果关系的显著性

在给定原因的情况下，某个结果出现的条件性概率也可以用来衡量原因的显著性。所以，我们可以观察休假能在多大程度上提高生产力这一结果出现的可能性。但是，很多不是原因的事件似乎也可以提升其他事件发生的可能性。如果工作时长和休假之间有一个共同的原因，那么它们看起来就像是彼此提高了对方出现的概率。

衡量原因强度的方法有很多，[24]但这些方法的基本理念都是要以某种

方式吸收其他信息来解释这些共同的原因。如果在休假和加班这两个变量都为真时，生产力提高的概率为 $X\%$，而只有加班这一个变量为真时，生产力提高的概率也为 $X\%$，那么知道休假信息并不能提升我们预测生产力提高这一事件出现概率的准确性。然而在实践中，我们可能也不会直接测量一个变量。也许我们并不能准确测量人们的工作时长，但我们知道他们在办公室待的时间有多久。有些人在办公室里可能会花很长时间吃午饭，也可能会花很长时间写私人邮件或者玩电子游戏。仅凭办公时间，我们无法将这些人和那些在办公室里待的时间较少但是工作效率更高的人区分开。因此，工作时长这个指标无法完美地将其结果进行区分。

这一点类似于我们前面看过的一些例子。在那些例子中，表示变量的方式（是将几个因素结合在一起还是一个一个单独研究）会影响推理的结果。所以，我们不仅需要一组变量来区分原因和结果，还应该认识到，由于有这些原因以及其他原因（数据缺失和测量失误等），没有因果关系的变量之间可能也会存在某种概率上的依赖性，我们必须想办法来解决这样的问题。

如果我们说休假可以提高生产力，那意思是休假或者不休假会对生产力产生影响。如果休假是一个十分重要的原因，而且不需要其他任何因素就能对生产力产生影响（比如需要足够的可支配收入以便休假不会造成经济压力），那么不管其他变量的值是什么（比如工作时间是长是短），休假之后的生产力都应该得到提高。然而，这一点并不是在所有情况下都成立。因为很多原因不仅能够带来积极的影响，还能带来消极的影响。比如说，安全带一般情况下可以避免交通事故中出现死亡事件，但在某些情况下，安全带却可能因为阻碍人们从落水的汽车中逃生而造成死亡事件。但我们仍然可以假设：即使安全带有时会导致死亡事件，但是系安全带死于交通事故的平均概率要低于不系安全带的平均概率。

因此，要想量化某个原因的显著性，我们可以计算这个原因平均在

多大程度上影响了其结果出现的概率。简单来说，就是在其他因素保持不变的情况下，这个原因出现和未出现时某个结果出现的概率会有多大的变化。可以将各种情况出现的概率进行加权计算。如果在一个非常普遍的情况下，一个原因可以显著地提高某个结果出现的概率，那么这个原因的显著性比那些只在极少数情况下才能提高某个结果出现概率的原因要大得多。

以图 6-8 中的因果结构为例。在这个图中，政党支持和意识形态会影响政客们的投票，选民的意见则不会。如果这组关系真是如此，那么不管选民是否支持这个法案，它获得投票的概率都是完全一样的。但是，如果意识形态或政党意见发生了变化，这个法案获得投票的概率则会发生改变。

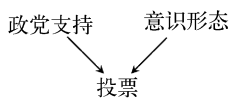


图 6-8　选民对投票的重要性的平均值会低至 0。注意，没有被圈出节点的图形不是贝叶斯网络

可以一次性确定所有变量的值，[25] 然后观察各种变量值的不同组合对结果产生的影响，以此来计算因果关系的显著性。一个政党可以支持或反对一项法案，意识形态可以符合或背离一项法案，选民也是如此。我们可以研究每一种可能出现的组合，然后观察在政党支持和意识形态这两个变量的每一种组合中，选民的支持会对结果产生什么样的影响。由于这两个变量完全决定了政客的投票，所以选民的支持不会对结果产生任何影响。然而，随着变量的增加，我们无法留意到每一种可能的情况，而且我们观察到的案例数量也不足以让我们得出任何具有统计意义的结论。为此，我设计了一个更实用的测量显著性的指标：在一个变量保持不变的情况下，不断改变原因的值（真或假），然后记录原因对结果的不同影响并计算它们的平均值。[26] 为了计算这个因果关系显著性指标的值（ε_{avg}），我们先算

出在政党对法案的支持保持不变的情况下，选民对选举结果的影响有多大，再算出在意识形态保持不变的情况下，选民对选举结果的影响有多大，以此类推，最后把这些不同的数值放在一起，求出选民显著性指标的平均值。

在大部分情况下，以上面这种概率为基础的计算方法都是从一个数据集出发，然后得出一个数字，这个数字会告诉我们一个变量作为原因对另一变量的显著性如何。这个显著性的值介于–1 和 1 之间，–1 表示的是一个非常强的、导致结果无法出现的负面原因，而 1 表示的是一个非常强的、一定会导致结果出现的正面原因。

由于在实践中总会出现一些噪声、失误和数据缺失的情况，所以我们不能假定不是原因的事物的显著性指标的值就一定为零。相反，我们经常需要确定哪些因果显著性指标的值具有统计意义（回忆一下第 3 章介绍的 P 值和多重假设检验）。[27] 比如说，有很多变量可能是某些变量的原因，但是它们之间又没有真正的因果关系，我们在计算这些变量的因果显著性指标的平均值时，会发现这些显著性值（ε_{avg} 值）的分布看起来就像一个钟形曲线，或者像图 6-9 中的浅灰色柱状图一样。而当测试的数据集中存在一些真正的因果关系时，这些显著性值会分布成另一种图形（图 6-9 中的黑色柱状图）。我们观察到的内容和期待观察到的内容之间存在一些差异，我们可以利用这种差异来判断显著性指标的哪些值应该被看成是具有因果关系的值。[28]

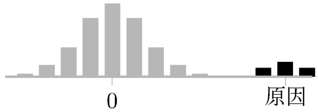


图 6-9　一组因果关系显著性值的柱状图。浅灰色区域（以 0 为中心，即不具显著性）代表虚假的因果关系，而黑色柱状图代表真实的因果关系。由于有噪声和其他因素，所以并非所有非原因变量的显著性值都是零，而是会以零为中心分布在邻近区域

要想保证因果显著性高的变量就是真正的原因变量，我们需要确保测量的因果关系强度是准确的(以便这些概率能够代表变量出现的真实概率)，还要确保也测量了变量之间的共同原因（否则我们可能会高估其他原因的显著性或者发现一些并不存在的关系)。对于时间序列来说，通常还需要假设这些关系不会随着时间的变化而发生改变。因为如果这些关系随着时间的变化而发生改变，那么可能就会出现这样的情况：两个变量在时间序列的一个时间段里是相互独立的，但在另一个时间段里却不然。在这种情况下，尽管两个变量之间的关系在一段时间内可能很强，但是当我们考察整个时间段时，两个变量之间的关系可能会显得很弱。

我们已经探讨了“为什么”的问题，但还没有讨论过“什么时候”的问题。在有些计算因果显著性的方法中，可以在原因和结果之间指定一个时间间隔或时窗，以便计算原因的显著性。如果与流感病人亲密接触后，接触者会在 1 到 4 天出现流感症状，那么这个时间条件就能让我们计算出二者之间的因果显著性值。但如果我们对引起流感的原因一无所知，怎样才能知道只要测试这个时窗就可以了呢？这些测量方法中的某些方法存在一个缺点，那就是如果测试的时间组不对，我们可能就会错过一些真正的原因，或者只能找到真正时间组的一个子集。我们不能为了解决这个问题而去测试每一个我们能想到的时间间隔，因为这样会大大增加计算的复杂性，而且这种做法也无法保证我们在实践中一定会找到正确的时间组合。因为这些数据样本并不是随着时间而均匀分布的，它们可能十分分散（数据量很少且间隔的时间很长)，并且数据之间的间隔可能也不是随机的。

假设我们有一组病人的一些实验检测结果以及他们的药物处方。即使某种药物会在一周之内导致病人血糖升高，那我们测试出来的数据也不会全部（甚至也不会是大多数）取自处方开出后整整一周的时候。此外，开处方的日期和服药的日期之间可能也会有一个间隔。所以从表面上来看，开处方和血糖升高之间的时间间隔好像延长了，但也许实际服药和血

糖升高之间的时间间隔真的只有一周。所以，我们在每一个时间间隔观察的数据可能还不够多。使用时窗有助于我们计算显著性值（因为如果将这些时间间隔统一放在一起，那我们大约观察 5 到 10 天可能就够了），但我们还需要搞清楚一个问题：到底要测试哪一个时窗？

要想从数据中找到时窗，可以先确定一个可能的时间段或者备选时间段，然后再根据数据调整这个时间段。显著性指标可以帮助我们实现这一点。在图 6-10 中，我们测试的时窗与真正的时窗重合了一部分，但也有不一样的地方，图中所列的就是各种可能出现的情况。随着时窗的放大、缩小和偏离，我们要重新计算因果显著性的值。在每一种情况下，把不正确的时窗改变得更加接近真实的时窗，显著性的值都会变大。有了时窗，结果变量本质上就成了在某个时间范围内出现的结果。如果测试的时窗比真实的时窗宽得多（如图 6-10 中的第一个长条形所示），那么就会出现很多这样的情况：我们很期待某个结果出现，但是这个结果却没有出现（由于在原因为真的情况下结果并未出现，所以这些案例会对原因的显著性值造成不利影响）。另一方面，如果测试的时窗太窄，那么即便测试的潜在原因没有出现，某个结果可能也会出现。随着测试时窗与真实时窗越来越接近，显著性值也会变大，并且最终会与真实的显著性值相一致。[29]

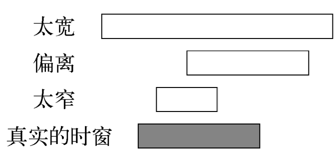


图 6-10　当一个原因的测试时窗与真实时窗有重合的地方也有不同的地方时，可能出现的各种情形

6.3.2　格兰杰因果关系检验法

概率主要用于包含离散事件的数据，例如已诊断或未诊断，或者被

划分为正常、偏高和偏低的化验值。但如果我们想知道一只股票的价格变化是如何导致另一只股票的交易量发生变化的，又该怎么办呢？其实我们真正想知道的不是一只股票的价格上涨会导致另一只股票的交易量上升，而是另一只股票的交易量预计会增加多少。概率法测试的是某个原因会导致某个事件出现的概率发生多大的变化，但我们也可以测试相对于原因发生的变化，某个变量的值会发生多大的变化。目前介绍的大部分方法都可以这样用。

严格来说，格兰杰因果关系[30]并不是传统意义上的因果关系（我们很快就会看到原因），但它却是在连续值时间序列数据中推理因果关系的常用方法。Wiener 曾经说过，原因提高了结果的可预测性。在 Wiener 的研究基础之上，格兰杰设计出了一个实用的方法，用于测试金融时间序列（比如股票收益率）中的因果关系。这个方法的基本思路是，原因给我们提供了一些其他变量没有的关于结果的信息，这些信息可以让我们更好地预测某个结果的值。所以，假设我们掌握了某个时刻之前的所有信息，在这种情况下，如果我们将原因从这些信息中剔除掉，那么结果为某个值的概率就会发生变化。

在实践中，我们掌握的信息并不是无限的，即便我们掌握了所有信息，也会由于计算的复杂性而无法将它们都派上用场。总的来说，有两种形式的格兰杰因果关系，每一种都能让你得出迥然不同的结论。值得注意的是，这两种形式的因果关系都不是真正的因果关系。但由于它们经常被用来证明那些因果主张，所以还是有必要了解一下它们能做什么以及不能做什么。

第一种形式的格兰杰因果关系叫作双变量格兰杰因果关系，这种因果关系比相关性强不了多少（它的值不是对称的）。它只有两个变量，而且只能告诉我们一个变量是否能帮助我们预测另一个变量。所以，如果我们在监测天气情况、航班晚点事件和机场咖啡的销量，只能发现两个变量

之间的关系，比如天气情况可以预测航班晚点事件。即便是在没有隐藏变量的情况下使用这个方法，也无法避免混乱的局面。所以，双变量格兰杰因果关系很容易让我们在由共同原因导致的各个结果之间发现并不存在的虚假因果关系。如果恶劣的天气导致飞机和火车都晚点了，那我们可能会错误地认为是飞机晚点导致了火车晚点，或者是火车晚点导致了飞机晚点。这种方法还可能会让人们认为，在一连串的原因中，是前几个原因导致了所有后来出现的原因，而不再是只发现各个原因之间直接的关系。也就是说，如果我们有一个事件序列，由于我们无法考虑序列中间的各个事件，所以可能会认为是第一个事件引起了最后一个事件。

用来测试格兰杰因果关系的方法有很多，回归分析就是其中一个简单的方法。假设我们想知道是先有鸡还是先有蛋。沿着 Thurman 和 Fisher 的思路，我们选取了两个时间序列，一个是每年鸡蛋的产量，另一个是每年鸡的数量。然后，我们会得出两个等式：一个表示鸡的数量如何取决于之前鸡和鸡蛋的数量，另一个表示的是鸡蛋的数量如何取决于之前鸡和鸡蛋的数量。“之前”具体是指之前几年（时间间隔或时窗），这个数字是由用户选择的一个参数。我们可以测算出某一年鸡蛋的产量和前一年（或者前两年，等等）鸡的数量之前的依赖程度，有一个系数可以告诉我们当前鸡和鸡蛋的数量与之前某一年鸡和鸡蛋数量之间的依赖程度有多强，系数为零意味着没有任何依赖关系。因此，在鸡蛋等式中，如果鸡的数量系数在某个时间段不是零，那么鸡和鸡蛋之间就存在格兰杰因果关系（如果之前某一年的系数刚好是 2，那么这就意味着当前鸡蛋的数量正好是之前某一年鸡的数量的两倍）。一般情况下，更多的时间间隔意味着更高的复杂程度，所以除了数据（比如数据点的数量和测量的粒度）上的局限性以外，在实际测量的内容上可能也有一定的局限性。

让我们再次回到机场。假设我们在预测咖啡销量时考虑了天气因素、航班晚点和之前的咖啡销量，这就成了多变量格兰杰因果关系。在这种因

果关系中，每一次测试都包含了所有的变量。尽管我们无法将世界上所有的信息都考虑进去，但是将所有的其他因素都考虑进去之后，就能测试出某个变量是否能提供一些有用的信息。假设真正的关系是这样的：天气恶劣导致航班晚点，航班晚点导致候机时间延长，而候机时间延长又导致咖啡销量上升。那么，一旦在咖啡等式中加入了航班晚点这个因素，天气情况就无法再为我们提供任何新的信息了，因此，它的系数应该接近于零（即它对预测咖啡销量不再会有任何帮助）。在实践中，我们不会只因为系数不为零就真的认为变量之间存在因果关系，而会做一些测试来看看这个不为零的系数在统计学上是否具有显著性。尽管多变量格兰杰因果关系更加接近因果关系，但我们无法保证发现的这些关系一定是真实的。更关键的是，尽管多变量格兰杰因果关系更有力也更准确，但由于它的计算强度太大，所以实际使用它的次数很少。[31]

6.4　现在该怎么办

或许你身上戴着活动监视器，搜集了好几个月的运动和睡眠数据；或许你从你们小区的报案记录中得到了一些数据，想从中找到犯罪的原因；或许你看到有人从社交媒体的帖子中发现了当地流感的流行趋势。那么，你该如何着手分析你搜集到的数据呢？

因果推理的方法不止一种，一定要认识到这一点。目前还没有哪一种方法能够在所有案例中都准确无误地找到事件之间的因果关系（这就让我们有了很多研究的机会）。有些方法得出的结论更具普遍性，但是这些结论取决于那些实际上不一定为真的假设。只知道一种寻找因果关系的方法并孜孜不倦地用它来解决每一个问题是不行的，我们需要的是一个工具箱。大部分方法都可以通过调整来适应大部分案例，但调整后的方法既不是最简便的，也不是最有效的。

没有一种方法是完美的，所以一定要了解每一种方法的局限性。比如说，如果你的推理是建立在双变量格兰杰因果关系基础之上的，那么你应该意识到，你找到的只是一种单向相关性，同时还应该考虑一下多变量的方法。如果因果结构（变量之间的联系）是已知的，而我们想要从一些数据中找出这个结构的各种参数（概率分布），这时贝叶斯网络也许是一个很好的选择。但是，如果时间是其中一个重要变量，那么使用动态贝叶斯网络或者研究因果关系时间变量的方法可能更合适。此外，我们研究的数据是离散的还是连续的也会限制我们所使用的方法，因为很多方法只能适用于其中一种类型的（而不是两者都适用）数据。如果数据中包含大量变量，或者我们并不需要找出完整的关系结构，那么用于计算因果关系强度的方法比推理因果模型的方法的效率要更高。但在使用这些方法时，还要考虑是否需要建立原因之间相互作用的模型，以便我们能够预测各种结果。因此，在决定使用哪些方法时，原因的用途和已有数据同样重要。最后还要认识到一点：在搜集和准备数据的过程中，我们所做的所有选择都会对最终推理出来的结论产生影响。

注释

1. FDA 不良事件上报系统（AERS）。
2. Tatonetti 等（2011）。
3. Tatonetti 等（2011）。
4. 一个重要的方法是快速因果推理法（通常缩写为 FCI）。想要了解更多详细信息，参见 Spirtes 等（2000）。也有一些研究曾将快速因果推理法进行扩展，以此来分析时间序列数据（Eichler，2010；Entner 和 Hoyer，2010）。
5. Meek（1995）；Spirtes（2005）。
6. 想要了解更多信息，参见 Andersen（2013）。
7. 人们除了尽量让数据稳定之外，还提出了几个专门用来推理非稳定性时间序列的方法。想要了解这方面的例子，参见 Grzegorczyk 和 Husmeier（2009）；Robinson 和 Hartemink（2010）。
8. 关于这方面的例子，参见 Pivovarov 和 Elhadad（2012）。

9. 关于这一研究的回顾，参见 Scheines（1997）。
10. 这个问题在哲学领域有一些争议。想要了解与这一看法对立的观点，参见 Cartwright（2001，2002）；Freedman 和 Humphreys（1999）。
11. 想要了解更多关于贝叶斯网络的信息，参见 Charniak（1991）。
12. 也就是 $P(B, A) = P(B \mid A) P(A)$ 。
13. 想要了解更多关于动态贝叶斯网络的信息，参见 Murphy（2002）。
14. 想要了解关于软件的综述，参见 Kevin Murphy 有关“Software Packages for Graphical Models”的内容。
15. Cooper 和 Herskovits（1992）描述了一个这种类型的早期方法。
16. 常见的就是贝叶斯信息标准（Schwarz，1978）。
17. Cooper（1999）。
18. 参见美国联邦储备委员会官网。
19. 随着变量数量的增加，可能出现图形的数量会呈超指数级增长。
20. Cooper 和 Herskovits（1992）。另一种办法就是定期使用一种新的、随机产生的图形重新进行检索。
21. 一种基于限制的方法就是 FCI（Spirtes 等，2000）。
22. 想要了解关于贝叶斯网络的更多内容，参见 Cooper（1999）；Spirtes 等（2000）。
23. Kleinberg（2012）。
24. 在 Fitelson 和 Hitchcock（2011）中有一个关于这一内容的研究综述。
25. Eells（1991）在处理原因显著性的平均度时正是这样做的。
26. Kleinberg（2012）的研究正是使用了这个方法。注意，在 Kleinberg（2012）的研究中，原因可能要比变量更为复杂，而且可能包含一段时间内的真实事件序列或属性。
27. 参见 Kleinberg（2012）第 4 章和第 6 章的内容，以获取更多关于计算原因显著性，以及如何选择临界值来确定某个数值在统计学上是否具有显著意义的内容。
28. 想要了解更多关于这方面的信息，参见 Kleinberg（2012）和 Efron（2010）。
29. 如何以数据推动的方式找到因果关系中的时间间隔？想要了解更多这方面的信息，参见 Kleinberg（2012）第 5 章。
30. 参见原文（Granger，1980）。
31. Barnett 和 Seth（2014）为我们提供了一个检验多变量格兰杰因果关系的工具箱。在很多平台中（包括 R 和 MATLAB 平台）都有双变量因果关系检验。

第 7 章　实验法

如何通过对人和系统进行干预来寻找原因？

与健康有关的很多说法似乎都经不起时间的检验，最终发生了逆转。最令人震惊的逆转之一是我们对激素替代疗法（HRT）和心脏病发作之间关系的认识：早期的研究发现，HRT 可以预防心脏病发作，但后来的研究却发现，HRT 对预防心脏病没有任何效果，甚至会提高心脏病发作的概率。

关于 HRT 好处的第一份证据取自护士健康研究（NHS）。[1]该研究调查了一个巨大的注册护士群体（近 122 000 人），产生了一定的影响。第一次护士健康调查发生在 1976 年，此后每两年对这些护士进行一次跟踪调查。人们在分析了 1997 年的调查数据之后发现，绝经后使用 HRT 的护士的死亡风险比其他护士要低 37%，而这主要是因为这个护士群体中死于冠心病的人数比其他护士群体要少得多。

后来出现了一些指导原则，表示可以使用 HRT 来降低患冠心病的风险。[2]但就在护士健康研究公布其发现的一年之后，就有人发表了另一项研究，声称 HRT 对冠心病没有任何疗效。与护士健康研究不同，心脏和雌激素/孕激素替代研究（HERS）试验[3]不是只观察人们的行为，而是将病人随机分成两组，一组使用 HRT，另一组只服用安慰剂。尽管这项研究只对 2763 名女性进行了为期四年的跟踪研究，却对护士健康研究的结论提

出了一些质疑，因为在研究的第一年，使用 HRT 的实验组的心脏病发病率不降反升（这一结果在最后两年被逆转了）。女性健康研究（WHI）的随机对照试验招募了一个更大的参与者群体，想要研究 HRT 对女性的长期影响，计划的平均跟踪研究周期为 8.5 年。虽然因为参与者群体中乳腺癌的发病率出现了显著增长，这项研究在平均研究周期达到 5.2 年后被中止了，但是研究人员发现了一个不可思议的现象：参与者的心脏病发病率提高了 29%（从每 1 万人 1 年中有 30 起心脏病发作事件增加到了 37 起）。[4]

HRT 似乎既可以加大女性心脏病发作的风险，又可以降低女性心脏病发作的风险，这是怎么回事呢？这是由不同的研究方式导致的。护士健康研究针对的是一个特殊群体，并且定时记录了 HRT 对参与者的影响、参与者所服用的药物以及其他情况。在这种观察性研究中，我们并不知道这些结果是否由某种药物导致，也不知道是否存在某个共同的原因导致参与者选择了某种治疗方案并出现了更好的结果——也许是由于护士们对自身健康的关心导致她们选择了 HRT 并且心脏病发作的风险也降低了。相反，随机试验可以排除病人特征和治疗方案之间的所有规律性。

干预措施常被视为检验因果推理活动的黄金准则。如果我们能够采取干预措施，并将参与者随机分到不同的实验小组中去（这里的实验小组既可能是接受治疗的病人，也可能是被指定要采用特定股票交易策略的股票交易人），那我们在一开始就能排除很多可能会导致人们选择某个干预措施或策略的干扰因素。然而，实际情况要复杂得多，因为我们并不总是能够采取干预措施，而且干预措施可能还会带来一些副作用，比如说服用降胆固醇药物的人可能不会那么注意自己的饮食。这一章将考察如何通过实验研究来帮助我们寻找原因，以及为什么一些声称找到了因果关系的研究可能无法重现，还会考察一个一般性的问题：为什么单独对一个事物采取干预措施如此困难？最后，我们还会讨论一些案例，看看为什么干预措施有时会让人们对事件背后的因果关系产生错误的认识。

7.1 从干预措施中获取原因

假设你想知道哪一种肥料能让你的植物长得最好。你先试了肥料 A，然后发现你的玫瑰没有开花。你又试了试肥料 B，然后你的花园突然充满了生机，于是你就确信这一切都归功于肥料 B——那种神奇的肥料。

这个方法的问题是什么呢？第一个问题在于你期望的结果——长得“最好”——是主观性的。也许肥料 B 的价格是肥料 A 的两倍，所以你想相信肥料 B 的效果比肥料 A 好，又或者你希望便宜的肥料和贵的肥料一样好。无论是哪一种情况，这些先验的信念都可能影响你对结果的判断（参见第 3 章介绍的证实性偏差）。

假设我们要通过量化评估来解决这些问题。我们可以数一数直径在两英寸以上的玫瑰花的数量并记录植株的高度。但是在这两个实验中，我们用的是花园中的同一块地，所以我们在使用肥料 B 时看到的变化可能是肥料 A 的延迟效应。这是在测试药物、饮食和其他干预措施的效果时经常需要考虑的因素。在交叉研究中，人们用同一个个体分别测试了 A 和 B 两种干预措施。在这个过程中，实验的顺序可能会影响实验结果，而且在评估 B 的效果时，可能还会出现 A 的剩余效应。比如说，服用营养补充剂后，该补充剂可能会在血液中停留一段时间。这时，我们需要在一个干预活动结束和另一个干预活动开始之间留有间隔，以便在评估第二个干预措施的效果时，可以排除第一个干预措施的遗留效应。最后，由于这两种肥料并不是同时测试的，所以也有可能在这两个时间段之间还有其他因素发生了改变。也许在测试第二种肥料时，雨水增多了或者日照时间变长了，所以植物的生长环境变得更好了。那么，这些植物的任何改善可能都是由使用肥料 A 和肥料 B 这两个时间段之间发生的变化引起的。

无论是用干预措施来比较不同的原因，还是用它来搞清楚某个事物到底是不是导致结果的原因，其实我们真正想知道的是，在所有其他变量

都保持不变的情况下，如果增加或者去掉某个可能的原因会出现什么结果。

从直觉上来讲，原因和干预措施之间存在一定的联系，因为我们经常把原因当作让事件发生的策略。而之所以想要找到具体的原因，是因为我们希望通过操纵原因来实现对结果的操控。使用观察数据来寻找原因的一个难题是，在有些情况下，可能很难将由共同原因导致两个结果的关系结构和由一连串原因构成的关系结构区分开来。比如说，在第一种情况下，某个政治候选人的演说可能会让他的人气更高并募得更多的竞选捐款；而在第二种情况下，竞选演说可能只会让他的人气更高，而人气的上升又导致他募得更多的捐款。如果我们能够独立操控捐款额和人气这两个变量，就很容易将这两种可能出现的因果结构区分开来。在第一种情况下，增加人气并不是获得更多竞选捐款的好办法（它们之间只有相关性），而在第二种情况下，增加人气可以让竞选者获得更多的竞选捐款（因为人气可以直接导致竞选捐款额的增长）。

由于存在这样一种联系，有些人曾试图从干预措施的角度来界定因果关系。大致来说，用正确的方式来改变原因会导致结果发生变化。[5] 当然，这种“正确的方式”还包括我们不会让产生某种结果的其他原因出现，也不会直接让结果本身出现。相反，我们想要确保只有原因才能对结果产生影响，以及干预措施不会对结果产生直接影响或者通过让其他原因起作用的方式以某种方法绕开原因。

假设演说、人气和捐款额之间的关系如图 7-1a 所示。为了测试虚线连接的两个变量之间是否存在真实的因果关系，我们可以通过干预措施来增加人气，以便观察人气上升是否会对捐款额产生影响。但这一干预措施可能也会导致知名度的上升，从而直接导致捐款额的上升，而不是通过影响人气而直接导致捐款额的上升。在图 7-1b 中，知名度直接影响了捐款额。同样，在图 7-1c 中，知名度通过让竞选者的演说机会增加而间接导致捐款额的上升。在第一种情况下，干预措施直接导致某种结果；但在第

二种情况下，干预措施导致另一个与干预目标不同的原因起了作用。这两种情况都存在这样一个问题：干预措施都是以另外一种方式导致某种结果出现的，而不是直接通过要测试的原因。

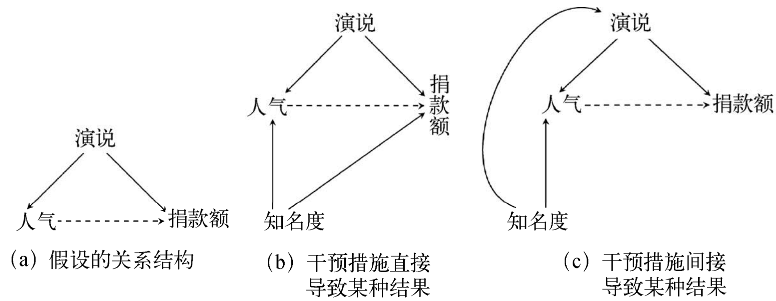


图 7-1　点线箭头标出的联系是被测试的关系结构。在其他图形中，实线箭头是起作用的箭头，虚线箭头是不起作用的箭头

7.2　随机对照试验

以上述那种理想化的方式操控一个变量是很难的。随机对照试验（RCT）解决了这个难题的一部分。在这种试验中，有两个或两个以上的小组，参与者被随机分配到各个小组中，所以不同小组之间的唯一差别应该就是处理方式的不同。因为所有其他特征的分布应该是一样的，所以如果出现了不同的结果，一定是处理方式不同导致的。这并不是一个"一键安装"式的理想化干预措施（比如增加钠元素的摄入量但不改变液体的摄入量），但是它却比其他任何措施都更加接近理想化。

但在使用 RCT 的结论时，这种严格的要求也是一种局限。在试验中，我们只考察一个变量；但是在现实生活中，试验结论却不一定能这样用。比如说，我们可能在 RCT 中发现某种药物对我们有益且没有任何副作用，但在现实生活中，人们在服用这种药物的同时可能还会经常服用另一种药

物，这两种药物之间会出现严重的相互作用，而这种相互作用可能只有在药物上市之后才会被发现。

虽然 RCT 经常出现在医药领域，但它们实际上只是一种实验研究方法而已，我们完全可以用这种方法来研究很多其他领域的问题。有一个著名的案例：谷歌使用用户点击数据来决定要在它的徽标中使用 41 种蓝色色度中的哪几个色度。[6] 通过将用户或访客随机分流到不同色度的或当前颜色的徽标的网页中去，对比不同网页中用户点击徽标的次数，从而测试出用户对不同色度的偏好。政治竞选也使用了随机对照试验来决定向选民传递什么样的信息以及如何传递这些信息。[7] 在政治竞选中，人们不再寻找投票行为和人口特征之间的相关性，也不再提出一些关于人们如何投票的理论，而是利用海量的电子邮件地址和详细的个人数据来测试各种干预措施的效果。比如，在某次竞选中，我们可以将具有某些特征的一群人随机分成不同的小组，给每个小组发送不同文本的电子邮件，或者给每个小组打不同内容的电话，以此来争取他们的捐款。试验的结果十分清晰（募得捐款的金额）。而且如果样本足够大的话，我们可以从很多不同的小组中测试出很多不同的信息。奥巴马在 2012 年竞选时正是这样做的，他们先在一个比较小的支持者群体中测试了不同特征的电子邮件取得的效果（例如不同邮件主题、不同建议捐款金额甚至不同邮件格式等），然后才把邮件发送给通讯录中的所有收件人。[8]

随着时间的推移，通过 RCT 获得的知识并不总是一成不变的（再次使用曾经有效的电子邮件还会有效吗），但是 RCT 已经被广泛应用于医药领域以外的很多领域（比如经济学和教育领域了）。即便你自己从来没有做过 RCT，能够评估试验的结果对于你的决策行为也是十分重要的。

7.2.1　为什么要做随机试验

18 世纪初，James Lind 做了一个试验并将其记录了下来，这个实验被

认为是第一个对照试验。试验发现，柑橘类水果可以很快治好坏血病。有一艘船上的很多水手都得了坏血病，James Lind 将症状相似的 12 名水手分成 6 个治疗小组。除了要测试的各种治疗方法以外，这 6 个小组的饮食是完全一样的。治疗方法包括醋、海水、柠檬和橘子。[9] Lind 发现，与其他小组相比，食用了柑橘类水果的治疗小组康复得很快。他由此认为柑橘类水果可以有效治疗坏血病。

然而，这个实验中的各个小组使用的治疗方案完全是由 Lind 安排的，而不是随机分配的。在他的记录中，采用海水治疗方案的那组水手的症状比其他小组要严重得多。[10] 尽管他的结论后来被证实是正确的，但如果我们根据症状的严重程度来选择治疗方案，那么这种区别性待遇完全有可能导致结果出现偏差（如果那些采用柑橘类水果治疗方案的参与者是一些症状较轻的、无论是否接受治疗都会康复的病人），或者导致与辛普森悖论类似的情形出现（如果那些采用柑橘类水果治疗方案的参与者是一些无可救药的病人）。要想避免在分配治疗方案的过程中出现区别性待遇，RCT 中的随机部分至关重要。

观察性研究的一个主要缺陷在于，人们关于是否采取行动和何时采取行动的选择会扰乱我们所观察到的事物之间的关系。比如说，我们很难测试玩暴力的电子游戏是否会导致暴力行为。因为我们并没有随机分配这些孩子去玩某种类型的游戏，所以即使二者之间存在某种相关性，我们也无法确定到底是电子游戏导致了暴力行为，还是暴力行为导致孩子们去玩暴力的游戏，又或者还有一个因素既导致了孩子们的暴力行为，又导致孩子们去玩暴力游戏。

同理，在护士健康研究中，护士们选择 HRT 的行为与她们面临患心脏病的风险以及她们对有益健康的行为的偏好是分不开的。也就是说，有可能 HRT 对心脏病根本没有任何效果，可能是那些选择 HRT 的护士们做了一些其他的事情，从而降低了心脏病发作的风险——而且有可能是她们

选择 HRT 的行为给我们提供了一些关于其他行为的信息，这让 HRT 预测到那些使用这一疗法的病人的身体状况会更好。还有一个类似的例子是，在所有其他药物都无效的情况下，用一些未标明用途的药物来对病人进行治疗。这就对病人采取了干预治疗，因此这些病人的治疗结果取决于他们疾病的严重程度、医疗护理的质量，等等。之前尝试的很多药物的剩余效应可能会进一步扰乱我们观察到的关系，让我们很难确定某种药物无效的含义到底是什么。随机试验的好处主要在于，它可以切断我们的选择（选择去干预的行为）和试验结果之间的联系。

假设我们将学校里所有 13 岁的孩子随机分成两组，然后给其中一组发信息，督促他们每天进行 30 分钟的体育活动，而给另一组发送的信息内容是天气预报。这两组孩子会互相联系，但我们无法知道他们是否会相互分享自己收到的信息内容，也无法知道那些收到督促信息的孩子是否会邀请他们的朋友（收到天气预报信息的孩子）一起参加体育活动。在药品的临床试验中，干预组的病人可能也会跟对照组的病人分享他们所服用的药物。这种小组之间的药物共享行为就是样本污染的一个例子。[11]

为了防止干预组和对照组之间出现样本污染问题，人们采用了一种群集设计，即随机分配各个群体而不是个体。比如说，我们不再对学生进行随机分配，而是将所有学校随机分成两个小组，给属于不同小组的学生发送不同的信息。针对药品临床试验的那个例子，我们不再对病人进行随机分组，而是将医院或医疗服务机构随机分成两个小组，每个小组中的病人接受不同的治疗方案。这种做法的前提是样本的规模足够大，这样我们才会对试验结果的准确性有同样的信心，因为同一个群集中的个体之间也许存在一定的相关性，而且各个群集的大小可能也不一样。这些群集可以是一个家庭（由于基因和环境因素，这个群集中的个体之间高度相关）或者一个学校（由于所处位置相同，这个群集中的个体之间仍然存在相关性，但是这种相关性要比家庭成员之间的相关性小得多）。[12]

无论是随机将个体分成两个小组还是将群体分成两个小组，这种将对象分成两个相似小组（除了对它们采取的干预措施不同以外，两个小组在各个方面都是相似的）的指令忽略了很多细节问题——它并没有告诉我们这两个小组（这两个小组不需要完全一样，只要它们之间具有可比性就可以了）中应该包括什么样的成员。因此，我们需要决定哪些人可以作为研究对象。

假设我们正在测试用于治疗胃灼热的药物。我们可以招募各个年龄段的不同性别的人参加测试，但是这些人中的很多人可能都没有患胃灼热。由于研究资金和研究周期都有限，这样招募参与者会浪费很多资源，而且大部分没有患胃灼热的人很可能也不愿意参加这样的测试。假设我们缩小参与者的招募范围，只招那些曾经得过胃灼热的人做参与者。这时，我们该不该招募那些由于另外一种情况（比如怀孕）导致出现胃灼热症状的参与者呢？我们是应该招募所有年龄段的参与者，还是应该只招募成年参与者呢？也许我们认为孩子们患胃灼热的生理过程与成年人有着本质的差别，因此决定只研究 21 岁到 65 岁有过胃灼热病史的参与者。然后又有了新的问题，这些成年参与者中可能有人已经在服药治疗胃灼热了，也可能他们身上还有其他可能会影响药效的因素。理想情况下，我们研究的参与者群体应该没有人吃过任何可能会和测试药物相互作用的药物。因此，我们可能会决定在 21 岁到 65 岁、有胃灼热病史且尚未每天服用治疗胃灼热的药物的参与者身上测试这种药物。

在选择研究对象的过程中，选择偏差可能会完全决定研究的结果。这种选择偏差可以是个体自主决定是否成为参与者导致的，也可以是其他因素让他们成为了或者未成为参与者导致的。我们在第 3 章讨论过，有些选择偏差可能会让人们去寻找那些偏向于特定结论的证据，也可能会影响我们评估搜集到的证据的方式。我们的研究方法可能也会导致数据出现各种

偏差。比如说，通过电话进行政治民意调查时，如果只调查有线电话用户而不调查手机用户，可能就会歪曲参与者的人口特征。例如在 2008 年，皮尤研究中心发现，在好几次只调查有线电话用户的民意调查中，奥巴马领先麦凯恩的百分比比实际上低了 2%~3%；而在大选之前的最后一次民意调查结果中，奥巴马领先麦凯恩的百分比比实际上低了 5%。[13]

随机试验的目的是限制选择偏差，但是在设计一项研究时，我们必须做出很多选择，而这意味着出现选择偏差的风险依然很大。参加一个实验是自愿的行为，所以选择参加实验的人和选择不参加实验的人可能在本质上就是两种人。如果一个登记参与者的研究人员知道了每一个参与者会被分到哪一组（如果分配规则就是参与者的登记顺序或者是一个登记人员已知的更为复杂的顺序），那这可能就会影响到研究者会把参与实验的机会给谁。这种偏差会直接影响我们能否通过研究得出因果结论（内部有效性），也会影响这项研究在参与者群体的典型度基础上的适用范围到底有多大（外部有效性，这一点将在本章后面讨论）。

接下来，我们看看如何处理那些参与者没有完成实验的情况。有些参与者可能会因为与实验无关的原因而中途退出实验，还有些参与者可能会因为无法接受干预措施（比如副作用超过了所有积极作用的干预措施）而中途退出实验。[14]在联系参与者获取实验结果数据时，有些参与者可能会联系不上（被称为“失访”参与者）。比如说，有一项研究要评估中风病人出院后 3 到 6 个月的恢复情况。研究方案可能会要求研究人员给这些病人或给他们的看护人员打电话并调查他们的恢复情况。但是，有些病人可能从来都不接电话，还有些病人可能已经换了电话号码或者已经搬家了，这就导致研究人员无法联系到他们。[15]

有些研究人员可能会在分析数据时直接忽略那些联系不上的病人，但如果这些数据的缺失不是随机的，那这样的行为可能会导致结果出现偏差。而且在评估一项研究时，缺失大量的参与者数据应该是一个很危险的

信号。假设我们要在老年人身上测试一种运动干预法。与不进行任何干预的对照组相比，那些每周运动 10 个小时的老人的胆固醇含量更低，而且寿命也比其他老人要长两年。然而，如果在随机分配去参加这项实验的老人中，有 75%的老人由于受伤或极度疲劳而中途退出了实验，那么这项研究很可能会得出这样一个结论：那些健康到每天可以运动一小时以上的老人比那些无法完成每天一小时以上运动量的老人活得更久一些。在这个案例中，老人是否一直留在这个实验中是评估干预措施的可接受性的关键因素。因此，直接忽略那些数据不完整的参与者会导致我们高估治疗方案的有效性并低估它可能会产生的副作用。

幸存者偏差可以被归为一种选择偏差，是由于我们在分析数据时只分析那些一直到某个时间点还幸存的或者还留在实验中的参与者。但更宽泛地讲，幸存者偏差是我们在分析研究结果时，只分析那些成功到达某个终点的参与者群体或案例群体的数据导致的。这个群体可能是那些至少有两年损益表的公司（忽略所有没到两年就破产的公司），也可能是那些已经完成第一任期的政客（忽略那些未完成第一任期就已经死亡的、辞职的或被提前赶下台的政客），还可能是那些已经发行过热门单曲的音乐人（忽略那些从没拿到过音乐发行合同的音乐人）。如果我们要研究的是频繁的巡回演出会给那些非常成功的音乐人带来什么影响，那么那些已经发行过热门单曲的音乐人可能就是我们要研究的对象；相反，如果我们要研究的是早期艺术教育对音乐上的成就会有什么影响，那么只研究这些在音乐上有很高成就的人就会让我们得出一个有偏差的结论。

在某些情况下，无论是出于道德考量还是成本因素，我们根本无法对参与者和各种情形进行随机处理，这时就需要利用其他类型的研究。一种是队列研究，比如护士健康研究。在队列研究中，我们会对一个群体进行一个前瞻性的、持续一段时间的跟踪研究。这种研究的缺陷（除了选择偏差外）在于，我们可以从每个人身上搜集同样的数据，但要想长时间跟

踪研究，那么研究成本会很高，而且中途退出的人数可能会很多。此外，倘若我们研究的结果出现的概率很小，那就需要一个很大的样本来进行研究，但这也无法保证我们能够观察到足够多的我们想要观察的事件。还有一种是病例对照研究，这种研究一般是回顾性研究。我们选择在某些特征上不同的（比如红头发的人和非红头发的人）两组参与者，然后回顾他们之间的差别是什么（比如遗传变异）。然而，由于我们只是在观察这些差异而不是积极地干预它们，所以无法保证已经测量了所有的干扰因子。

7.2.2　如何设置对照组

1946 年，Bradford Hill 和英国医学研究理事会的其他研究人员一起对比了卧床休息和服用抗生素链霉素对治疗肺结核的效果。这是医学研究史上具有里程碑意义的一件大事，而且可能也是医学研究中第一次使用随机对照试验。[16]参与这次试验的每一家医院都收到了一组标了号的密封信函，每一封信中都装了一份治疗方案（休息或链霉素）。由于医院里的每一个肺结核病人都登记参加了这次试验，所以与每一个病人的登记号码的下一个号码相对应的信封就是这个病人的治疗方案。[17]

与 Lind 的试验一样，研究人员并不只是考察服用链霉素前后的效果，而是将链霉素的效果和当时标准的治疗方案（即卧床休息）进行了对比。这一点很重要，因为只对比病人在接受治疗前后的病情的话，即使治疗方案没有任何效果，如果病情本身随着时间的变化逐渐好转了，那么对比结果也会显示病情有所好转；而且在某些情况下，治疗行为本身（即便这个药物没有效果）可能也会给病人的病情带来积极的影响。

比如说，有些病人深信某种抗生素可以治疗他们身上的流感病毒，有时他们会一直要求医生给他们开这种药，直到医生同意为止。如果最后他们的感冒跟大部分人一样都好了，那他们的康复和这个药也没有任何关系，这只不过是感冒过程的必然结果。如果他们在生病时没有服用抗生素，

而是喝了杯咖啡、看了很久的电视，或者做了任何其他的事情，那么这些行为似乎对治疗感冒也有同样的效果。

进行对照试验的另一个原因是，我们在实际生活中并不是在新的治疗方案和没有治疗方案之间进行选择，而是想知道一组治疗方案中的哪一个效果最好。在选择合适的对照组时，由于我们不应该妨碍病人接受有效的治疗，所以不仅需要考虑道德问题和逻辑问题，还必须解释治疗行为本身对结果的影响。

在某些情况下，我们可以将标准治疗方案的治疗效果和新方案的治疗效果进行对比；另外一些情况下，我们可能需要安慰剂。这可能是因为没有标准治疗方法可以参照，也可能是由于研究方法中存在偏差。即便是使用一种比当前使用的治疗方法糟糕得多的治疗方法，可能也比不接受任何治疗的效果要好。选择一种合适的安慰剂很难，但从根本上来说，安慰剂是一个能够尽可能模仿真实干预措施而又不具备真实干预措施的主要有效特征的事物。最简单的例子就是，如果某种药物是以药丸的形式服用的，那人们通常会用一个跟这个药丸一样的糖丸来作为安慰剂。如果干预措施是有关改善健康的短信，那人们可能会用与健康无关的短信来作为安慰剂。不过，要想找到一个用来代替针灸的安慰剂就难得多了。在最极端的情况下，人们甚至曾经在帕金森和其他疾病的治疗试验中，使用虚假的手术来解释手术行为本身对病人的影响。[18]

一种治疗方法虽然没有任何已知的有效成分，却仍然能够改善病人的病情，这就是安慰剂效应。它能导致一些很奇怪的结果，[19] 甚至在病人已经知道他们吃的是安慰剂时，仍然可能会出现安慰剂效应。[20] 有报告指出，有些病人在服用安慰剂时出现了副作用。[21] 而且在安慰剂对比研究中，人们还发现由于药丸的剂量（似乎药丸数量越多效果就越好）和外观不同，安慰剂的治疗效果（安慰剂效应）也不同。[22]

这让我们想到了链霉素试验的另一个关键特征：这是一个双盲试

验，无论是参与试验的病人还是评估治疗效果的研究人员，都不知道病人接受的是哪一种治疗方案。[23] 这是避免证实性偏差的关键步骤——因为那些预计某种药物会生效的病人，可能会以另外一种方式向医生描述他们所出现的症状；同样，如果医生知道病人接受的是哪一种治疗方案，他们可能也会对病人的病情做出不同的诊断。

有一项研究在测试多发性硬化症的多种治疗方案的同时，还测试了盲法试验对试验结果的影响。在实验中，两组不同的神经学家对同一群病人进行了评估，其中一组神经学家对病人所接受的治疗方案一无所知，而另一组神经学家则清楚地知道每一个病人所接受的治疗方案是什么。经过 24 个月的定期观察后，接受盲法试验的神经学家们发现没有一个治疗方案是有效的，[24] 而那些没有接受盲法试验的神经学家却发现有一组病人的病情有所改善。之所以会出现这个差别，是因为神经学家对病人病情的评估是定性评估，所以那些没有接受盲法试验的神经学家可能因为已知病人接受的是何种治疗方案，因此在评估病人的病情时受到了影响。当一个实验的结果涉及这样的知识时（无论是评估对照试验中病人的病情，还是评估自家花园里花儿的生长情况），知道每一组参与者接受的干预措施是什么，可能会改变我们对已有证据的解释方式。

一般来说，在单盲试验中，病人不知道他们接受的治疗方案是什么，但是做试验的那些人是知道的。在双盲试验中，无论是病人还是临床医生都不知道病人接受的治疗方案是什么。然而，当搜集完试验中的所有数据后，我们不能简单地将这些数据都放进一个黑匣子，然后等它给我们输出一个确定的结果。在数据分析中，我们需要做很多决定（比如要做哪些统计检验），而这些决定可能也会存在偏差。因此，还有一种选择就是做三盲试验。三盲试验首先是双盲试验，同时，那些负责分析数据的研究人员也不知道每一组参与者接受了什么干预措施。[25]

这种三盲试验也许并不总是行得通，但我们可以在看到任何搜集到

的数据之前，预先确定分析数据时将会采取的所有步骤，并把这些步骤记录下来，以表明数据分析方案是在没有受到结果影响的情况下独立设计出来的。[26]实验和药品试验登记正是这样做的，这要求研究人员在搜集到任何数据之前就确定数据分析方案。[27]由于经常会有意想不到的情况出现（虽然这会导致人们倾向于发表积极成果的偏差很明显[28]），所以这种方法在实际中也会遇到一些问题。在我们假设的胃灼热研究中，我们可以提前确定要测量的主要指标（比如胃灼热出现的次数）和次要指标（比如胃灼热的严重程度）是什么，并且提前确定如何进行盲法试验以及大约会有多少参与者参加这个试验。然而，我们招募到的参与者数量可能达不到目标，也有可能我们没有预计到会资金缺乏而不得不提前结束试验。所以，严格按照预先制订好的方案执行也许并不总是行得通。

7.2.3 研究结果适用于哪些人

假设我们确实进行了胃灼热研究，而且研究似乎很成功。与另外一种治疗方案相比，这种药物大大降低了胃灼热的严重程度并减少了病人的发病次数。于是，这种药物最终获得了生产许可并开始投放市场。一位医生曾经看过这个试验的结果，他现在在诊断一个 80 岁的病人，这个病人每天服用 10 种药物，[29]而且不仅患有糖尿病，还有充血性心力衰竭病史。这位医生该不该给这位病人开这种新药呢？

通过控制试验来保证内部有效性（这意味着它能够回答我们提的问题）常常会牺牲外部有效性（试验结果具有更广泛的普遍性）。研究一个同质群体可以将可能的原因分离出来，但当我们在做关于其他群体的决定时，这反过来又会限制这种研究结果的适用范围。另一方面，研究群体中更多的变化可能会引起混乱，并导致我们无法找到变量的真正影响（如果这种影响只在特定子群中出现的话）。因此，我们必须认识到，随机对照试验的每一个阶段都有一个选择的过程。

在一个典型的临床试验中，我们所用的潜在患者库是那些在进行临床试验的机构接受治疗的患者，或者是那些机构能接触到的患者。但是，这个患者库已经将所有无法获得或者没有寻求医疗服务的患者排除在外了。我们需要想一想，临床试验机构和参与试验的临床医生治疗的都是什么样的病人。这些病人可能会比整个病人群体的平均病情更加严重，或者正好相反，这些病人中可能并不包括那些病情最严重的病人，因为他们已经被介绍到其他地方接受治疗了。我们还需要考虑这些试验所设定的病人的资格标准，这些标准常常将那些同时患有多种慢性疾病的病人排除在外（正如我们在假想的胃灼热试验中的做法）。等到病人真正同意参与这个试验时，试验针对的患者群体已经被筛减掉很多人了。这个试验其实并不一定要包含所有的患者，重点是试验中存在很多实际考量的因素，它们会影响最终哪些人会被招募进来。当不再考察试验的有效性而是开始试图应用试验结果时，我们需要考虑到这些因素。

如何确定试验结果是否适用于某个病人或病人群体？人们已经对这个问题进行了很多讨论。[30] 我们通常并不是在 RCT 的理想化世界中做决策，比如病人通常只患一种疾病。在大多数情况下，我们也不可能等到有了正好相关的研究之后再做决策。临床医师在为病人确定治疗方案时就是如此，我们在试图确定研究报告和自身的相关性时也是如此。RCT 的问题在于，它只告诉我们某种治疗方案在某个特定的人群中可能会导致某种结果。但是，另外一个人群却可能不具备让这种治疗方案起作用的特征。

比如说，如果一个 RCT 发现药品 A 比药品 B 更有效，而另一个 RCT 发现药品 B 比药品 C 更有效，那么我们很可能会假设药品 A 比药品 C 更有效。对各种抗精神病药物的考察发现，情况正是这样的。但是，这种随机试验还证明了药品 C 实际上要比药品 A 更有效。[31] 为何会出现这种反常的发现呢？很多这样的研究都是由被测试药品的制造商资助的，但无论这些研究是由谁资助的，这种不一致的结论都会出现。即便报告的数据完全

真实，并且在试验的过程中没有任何不道德的行为，但由于人们必须在试验的过程中做出各种决定，所以实验的结论仍有可能会偏向某种结果。通过选择特定的服用剂量、资格标准、结果指标以及统计检验指标，每一个选择都可能会偏向某种药物，从而导致这种药物的效果看上去比其他药物更好。

我们将在第 9 章学到，要想真正确保试验的结果也适用于一个新的群体，我们需要确保让原因有效的这些特征同时存在于试验群体和新群体中，还要确保新的群体中没有任何会对原因产生负面干扰的特征。然而，这对我们来说是一个巨大的负担，因为我们往往并不知道让原因起作用的因素是什么。假设我们随机给人们分配不同类型的办公椅，以便观察与坐在普通办公椅上相比，坐在瑞士球（抗力球）上是否会导致人们的体重下降。在实验中，坐瑞士球使人们在 6 个月内体重有所下降具有统计学上的显著性。但是，当我们在一个新的群体中做试验时，坐瑞士球却没有对他们产生任何影响。如果情况是这样的：第一组参与者发现瑞士球坐着很不舒服，或者他们经常会从瑞士球上摔下来，所以他们每天上班的时候会站很长时间，或者一直来回走动；而第二组参与者把瑞士球当作椅子一样，坐在上面就不再到处走动。在这种情况下是会出现这种结果的。其实真正的干预措施并不是瑞士球，而是一个让人们多起身、多走动的东西，但我们在这个研究中并不一定能看到这一点。同样，一种干预措施在受控环境下的使用方式也许并不能反映它的真实效果。有些药品需要在每天的同一时间准时服用，如果试验中的病人遵从这一医嘱的可能性更大的话，那么这种药品在现实中表现出的疗效可能会比在试验中差一些。

一项研究结论的使用方法会受到很多其他因素的影响，比如跟踪研究周期的长度。如果某种新治疗方案的 RCT 只持续了很短一段时间，那么我们可能就会怀疑长期使用这种治疗方法是否会有同样的效果，以及是否存在一些只有在服用了很多年之后才会出现的副作用。研究周期还可以

决定内部有效性。如果一项研究测试的是服药提醒短信是否可以提高病人遵从医嘱的概率，但这项研究只跟踪研究了病人三天，它就不能有效地证明短信提醒一般可以在很长一段时间内提高病人遵从医嘱的概率，因为人们对新的干预措施的热情往往会随着时间的流逝而减退。尽管如此，由于成本的制约，我们不得不在跟踪研究周期长度和样本规模之间进行权衡。

人们设计了一些检查清单和指导原则来评估 RCT 的研究结论，并规定了在一项研究中应该报告哪些内容。[32]需要注意的一点是：我们不仅要考察一项研究的内部有效性，还要考察它的外部有效性。这二者各自的重要性有多大，则取决于我们的目的是什么。有些内部有效性比较低的研究可能会通过较高的外部有效性（而且可能会与我们关注的人群有更多相关性）使其重要性得到增强。[33]在评估的过程中，我们主要需要关注的问题有：都研究了哪些人？研究的病人是如何挑选出来的？研究是在哪做的？跟踪研究的周期有多长？对照组是什么？这项研究是如何进行盲法设计的？

7.3　当参与者只有你自己时应该怎么办

在很多情况下，我们要做的并不是判断要给某个人群推荐哪一种药物，或者哪一种饮食方案最好，而是要做一些关于自己的决定。哪一种药物能更有效地减轻我的头痛症状？长跑之后要洗冷水澡还是热水澡才能让我的体力恢复得更快？早上喝几杯咖啡最好？

我们通常并不会系统地处理这些问题。相反，决定服用哪一种抗过敏药的过程更像是一个反复试错的过程。首先，你可能会去看医生，然后医生会给你开一种抗过敏药。服药一段时间后，你可能会胃不舒服，于是又去看医生。也许医生会调整你的服药剂量，但调整剂量之后你的过敏症状又复发了。于是，你再次去看医生，询问医生可不可以试试另一种抗过敏药。你可能会按照医生嘱咐的疗程去服用第二种抗过敏药，也可能会因

为过敏症状似乎好多了就提前停止用药。下一次你去看医生的时候，医生会问你第二种药的效果怎么样，而你并没有觉得这个药有什么问题，所以你会说这个药很有效。这是否意味着你从一开始就应该服用第二种抗过敏药呢？

从根本上来说，这和我们在本章前面讨论过的肥料试验面临同一个问题。这种不系统的序贯试验不仅不能告诉我们这两种治疗方案哪一个更有效，甚至也不能告诉我们哪一个对你来说更有效。然而，由于试验中只有一个参与者，我们自然无法进行 RCT——随机让一个人接受要测试的治疗方法，让另一个接受与之进行对照的治疗方法。

与随机挑选病人不同，只有一个参与者的试验（被称为 n-of-1 trial，即基于单个患者进行多重交叉设计的随机对照试验）随机安排治疗方法的实施顺序。[34] 在这方面，前面介绍的肥料试验尤其缺乏说服力，因为我们只测试了一个序列（A-B），而且不知道在测试肥料 B 时肥料 A 是否依然起作用，也不知道在测试肥料 B 时的环境是否恰巧对花儿的生长更有利。每一种干预方法只测试一次并不能得出非常有把握的结论，所以一般情况下会重复测试多次。但是，要确定这些干预措施的实施顺序却有点复杂。我们似乎只需要重复测试 A-B 序列就能获得更多的数据，比如测试 A-B-A-B。尽管对于每一种干预措施来说，我们都有了双倍的数据，但是这两种干预措施的顺序并未发生改变，B 总是在 A 的后面。如果我们测试出的结果指标随着时间而发生了缓慢的增长，那么即便干预措施是一样的，B 的效果似乎也总比 A 要好，因为我们衡量 B 效果的时间要比 A 稍微晚一些。此外，如果我们采用的是盲法试验，那么这种简单的轮流法可能会导致个体能够猜出干预措施的安排顺序。

从理论上来说，人们可以在每个时间段随机选择两种干预措施中的任意一种，但是这种策略也存在一些问题。因为这种做法无法保证每一种干预措施被选中的次数是一样的，也无法保证两种干预措施的分布是均匀

的，所以有可能会出现一个全部由 A 组成的序列后面跟着一个全部由 B 组成的序列。这种做法一方面会导致结果出现偏差，另一方面，如果试验还没有测试到 B 序列就已经被提前终止了,那么这会将这项研究置于十分尴尬的境地。因此，我们可以将两种干预措施组成一组，然后随机安排每一组的实施顺序。这样一来，一旦选择了 A 干预措施就意味着接下来要实施 B 干预措施。但是，这样做仍有可能会导致轮流序列，所以我们还可以采用另一种方法，在 A-B 和 B-A 序列之间寻求平衡，每个 A-B 序列之后跟一个 B-A 序列，或者每个 B-A 序列之后跟一个 A-B 序列。也就是说，第一组要么选择 A-B 序列，要么选择 B-A 序列，然后没有选上的那一组就作为下一组。于是，有可能会出现这样一个序列：B-A-A-B-A-B-B-A。回顾一下前面讨论过的非稳定性问题（见第 4 章），我们现在要做的就是削弱这些时间趋势对结果的影响，以及削弱干预措施的实施顺序对结果的影响。

假设我们现在确定了一个用来测试两种干预措施的序列，但是第一种干预措施的影响会持续很久。那么，干预措施 B 很可能会得益于干预措施 A 的影响。在一个标准的 RCT 中，每一个参与者只接受一种干预措施，所以我们无须担心多种干预措施带来的累积效应，或者多种干预措施之间的相互作用。然而，在序列试验中，不仅实施的顺序可能会影响干预措施的效果（比如在对两种界面进行测试时，人们可能总是更喜欢第二种界面），而且每一种干预措施都可能会有一些持续很久的影响（比如人们对系统越来越有经验，而这可能会提升干预措施的效果）。在试验肥料的案例中，如果肥料 A 起作用的速度比较慢，不过一旦生效就会产生比较持久的影响，那么肥料 A 产生影响的时间段和使用肥料 B 的时间以及测量肥料 B 效果的时间就会有重叠。解决这个问题的办法之一就是，在使用了肥料 A 之后，过一段时间再开始使用肥料 B。这就是清除期，它的目的在于保证第二种干预措施实施的时候，第一种干预措施产生的任何影响都应该已经消失了。但一种药物的积极效应可能会很快消失，它的副作用却会存

留很长时间。清除期的另一个局限性是，它要求我们在一段时间内不能采取任何干预措施，但在一段时间内不采取任何治疗方法可能并不是人们想要的状态（比如在测试止疼药的时候）。要想确定一个合适的时间段作为清除期，还需要我们对干预措施的工作原理有足够的背景知识。解决上述问题的另一个方法，就是连续实施这些干预措施，但是忽略每一种干预措施一开始时搜集到的部分数据。

这种试验（针对一个参与者进行的 RCT）要求研究对象不会随着时间的变化而迅速发生改变，所以它的适用范围有限。对于像流感这样的急性病来说，以一个病人作为参与者进行试验没有任何意义。但对于像关节炎这样的慢性病来说，人们就成功地使用了这种试验。[35]同样，与选举这样的一次性事件（因为在选举之前的几周内，很多事情都在不断变化）有关的序列试验也没有任何意义。这种试验的最佳研究对象是那些或多或少具有一定稳定性的事物。

7.4 可再现性

在一项研究中，我们使用了一组电子病历来分析引发充血性心力衰竭的危险因素。研究发现，糖尿病是一个会引发充血性心力衰竭的危险因素。但当我们使用另一群人的病历数据来复制这个研究时，却发现充血性心力衰竭和糖尿病之间没有任何联系。相反，我们发现医生给病人开的胰岛素倒成了引发充血性心力衰竭的危险因素。[36]我们应该如何解释这两种充满分歧的结论呢？

要想复制一项研究，最好在完全一样的情况下使用完全一样的研究方法，这对于确保研究方法的说服力以及研究结果的可靠性至关重要。值得注意的是，复制研究和再现研究是不同的，后者的目标是要引入变化来测试研究结论的普遍性。复制研究则包括共享计算机编码、原始数据以

及执行计算机编码所需的所有步骤。如果其他人能够从中得出完全一样的结论，那么这个研究就是可以复制的。在有些实验中，一些细微的变化都可能导致结果出现巨大的差异，所以真正意义上的可复制性研究很难实现。即使是在计算机程序（似乎每一次执行这个程序的时候，它都应该有同样的表现）这种案例中，一个隐藏的病毒可能也会导致程序出现无法预测的行为。

当我们在科学研究中谈论可复制性时，通常指的是可再现性。也就是说，我们想要知道，在一项研究中发现的结论能否被另一些研究人员在另一个情况稍微有所变化的研究中再次发现。[37]这能够更加有力地证明我们发现的结论并不是巧合。假设有一项研究发现，与得到胡萝卜相比，孩子们在得到一块两盎司的巧克力之后，心情改善的程度要大得多。这项研究的主要发现是，和蔬菜相比，给孩子们巧克力会让他们更高兴。所以，另一项研究可能会通过改用 M&Ms 巧克力豆和西兰花再现这一结论，还可能会改用好时之吻巧克力和红薯再现这一结论。这些研究都没有复制最初的那项研究，但是它们都再现了最初那项研究所发现的结论（与蔬菜相比，巧克力让孩子们更开心）。

再现研究结果对于观察性研究来说格外重要（如果无法再现研究结论，可能说明研究中还有未考察到的共同原因），而再现实验研究中发现的结论对于形成普遍的认知来说也至关重要。此外，鉴于我们在实验中需要做很多决定，如果我们无法再现研究结论的话，可能就会发现很多研究会出现的各种偏差甚至是不当行为。

有些研究的主要发现无法再现，最近的很多研究工作也十分关注这一情况。医药公司的一些报告暗示，从科技论文中找到的药物靶标只有20%~25%是可以再现的。[38]还有一项研究发现，在 53 项关于癌症的主要研究中，只有 11%的结论是可以再现的，[39]而一些观察研究中的样本再现率则更加糟糕。[40]人们还试图再现那些备受瞩目的心理学研究结论（因为

这些结论常常会成为很多其他研究者的工作基础），但结果却一言难尽。[41]

为什么在一项研究中发现的真实的因果关系在另一项研究中却可能发现不了？

这可能是虚假数据或意外失误（比如电子表格中的打印错误或实验室污染[42]）之类的问题导致的。除此以外，真实关系的再现也并不像看起来那么简单。在研究心力衰竭的案例中，我们确实再现了研究结论，但要想真正弄清楚这个问题，还需要掌握很多关于变量含义的背景知识。当将糖尿病的诊断记录以一种结构化的格式和诊断时间保存在一起时，我们发现了它们和充血性心力衰竭之间的联系。但在第二个人群中，我们却发现胰岛素（治疗糖尿病的一种药物）成了导致充血性心力衰竭的一个原因。因为药物是以结构化形式保存的为数不多的事物之一，所以药物出现的时间或者存在与否都是更加确定的信息。另一方面，由于这种医学研究使用的是医院的病历，所以我们甚至无法确定谁得了什么病。此外，我们不一定总是能在不一样的地方搜集到完全一样的数据。

假设我们并没有再现研究结论，那是否就意味着最初的研究结论是假阳性结论，还是最初结论的普遍性只是比我们想象得小一些而已？也可能我们本就不应该指望在当前测试的这个人群中再现这一结论。比如说，有研究发现人们在因果判断中存在文化差异，所以某种因素在某个地方可能确实会影响人们的因果判断，即使这个结果在另一个地方无法再现（即这个因素在另一个地方不会影响人们的因果判断），它在这个地方也是真实的。这并不是说哪一个研究结论是错误的，而是说这个发现可能是最初研究的人群所特有的，或者最初研究的人群可能具有某个我们不一定知道的特征，而我们的发现正是这个特征所特有的。在这个案例中，我们尝试进行的复制研究是有价值的，它告诉我们这一发现在什么时候会出现，什么时候不会出现。

也可能会出现这样的情况：当我们发现这个因果关系时，它确实是

存在的，但是后来再对它进行测试的时候，由于人们已经知道了这一因果关系，所以整个系统发生了改变。比如在金融领域，人们发现了某种可能会影响交易行为的因果关系。[43]这种因果关系在研究期间可能是真实的，但它是不可复制的，因为这种因果关系随着时间的变化已经不再真实了，或者是因为我们利用这种因果关系实施了干预措施，然后改变了人们的行为（更多内容见第 9 章）。随着人们对广告信息敏感度的降低，以及竞选对手通过投放广告来进行针锋相对的回应，电视广告对政治候选人的积极影响和消极影响可能也会减弱。但如果某项研究试图去推断一般性的人类行为，且它的观点超越了研究对象和研究周期的限制，那倘若我们未能再现这项研究结果，一般而言，这个观点就已经被推翻了。

当然，在很多情况下，倘若我们无法再现这些研究的结果，可能就意味着最初发现的那个关系是假的。这些结果也许是人们使用的研究方法人为导致的，也可能是分析过程中的错误导致的，还可能是研究方式上的偏差导致的。很多影响外部有效性的因素同样会影响可再现性。我们在第 3 章介绍过一个死三文鱼的实验，那个实验得出错误结论的原因是试验的次数太多了。虽然我们通过修正对比分析的次数最终解决了这个问题，但如果我们发现的结果只是噪声的话，那么那些使用一条（或两条）新的三文鱼进行验证的试验，应该会在三文鱼的大脑中发现不同的活跃区域。

7.5　机制

如果有人告诉你海盗让地球的平均气温下降了，你肯定会觉得这是不可能的事。但是你的怀疑并不是通过实验而获得的结论，你没有在改变海盗数量的情况下观察地球的气温有没有发生变化，也没有证实海盗数量和地球气温这两个变量之间是没有相关性的。相反，你之所以会排除这样的可能，是因为根据你对这个世界运行机制的了解，你无法想象出有哪一

种方法能够通过改变海盗的数量来改变地球的温度。同理，我们之所以会认为有些因果关系是可能的，也正是基于我们对这个世界运行机制的认知。由于我们已经掌握了紫外线照射和皮肤癌之间的联系，所以即便没有任何观察数据，我们依然可以预测室内晒黑床和皮肤癌之间可能存在一定的联系。

这是一种关于事物运行机制的知识，或者说是关于某个原因如何导致某种结果的知识。虽然在不知道事物运行机制的情况下，我们也能找到导致某些结果的原因，但是运行机制是能够证实这种因果关系的一个证据，而且我们能够通过它找到更好的干预措施。原因会告诉我们某些结果为什么会出现，而运行机制会告诉我们这些结果是如何出现的。比较一下“吸烟让手指变黄”和“香烟烟雾中的焦油给手指上的皮肤染了色”这两句话。有些研究曾试图用事物的运行机制来定义因果关系。在这些研究中，运行机制大约相当于一个系统，其各个组成部分相互作用，通常会让事物发生某些变化。[44]不过更重要的是，运行机制可以为我们提供一些用来寻找因果关系的线索。

到目前为止，无论是使用经常出现的模式、概率的变化，还是剂量反应关系，我们所考察的各种证明因果关系的证据都与原因和结果一起出现的频率有关。如果几经观察，我们发现得了流感的人出现发热症状的概率更大，那么就会由此得出这样的结论：得流感会导致发热症状。但是，我们也可以根据事物的运行机制推理出这样的结论：身体中出现的感染病毒会给（控制体温的）大脑发送信号，然后大脑会将体温上升，以此来抑制感染症状。一部分信息告诉我们这个原因是如何导致某种结果可能出现的，另一部分信息则告诉我们这种结果实际上是会出现的。[45]

不过，正是由于对事物运行机制的认识，只用两个基因变体来解释选民投票率这种复杂的现象是不合理的。如果这些基因同时还与很多疾病以及其他现象有所关联，那情况就更是如此了。[46]就运行机制而言，似乎

不可能存在某种既能让人们更有可能去投票，又能导致肠易激综合征这类疾病的机制。情况更有可能是这样的：这两种结果都涉及很多因素，而我们发现的基因也许只不过是事物变化过程中的一部分而已。

有人认为每天刚好喝两杯咖啡对健康有益，这种说法似乎也不可信。因为我们很难想象有哪一种机制能让两杯咖啡对健康有益，而不是一杯半或者两杯半。因此，即使某项研究表明喝这么多咖啡和对健康有益这两件事在统计学上具有显著性，我们仍有可能认为一定还有某个其他原因导致了这样的结果。相反，人们在看到剂量反应甚至 J 形曲线（就像我们在第 5 章看到的那样）时，似乎都没有如此意外。这是因为很多生理过程都会导致这一结果，而只服用一剂药就会产生效果的生理过程则要少得多。

只要有人提出了某种运行机制，有人就可能去做一些揭开事物之间因果关系的实验。比如说，如果我们不知道导致某种疾病的原因是什么，但是知道有一种机制可能会导致这种疾病，而且还有一种针对这种疾病的药物，那么观察这种药物是否有效可能会为我们提供一些关于原因的线索。关于运行机制的知识还有助于我们设计出更好的干预措施。如果我们只知道感染疟疾的蚊子会传播疟疾，却不知道疟疾是如何通过蚊子传播的，那么防止疟疾的唯一措施可能就是阻隔人与蚊子之间的接触。但是，如果我们知道疟原虫进入血液之后会出现什么情况，那就能够获得多种控制疟疾的方法，比如阻止疟原虫进入肝脏，以及阻止疟原虫的繁殖，等等。

7.6　实验法是否足以找到事件发生的原因

虽然实验法和 RCT 能够在很多方面给我们提供帮助，但我们有时却不能或者不应该使用这些方法。我们不需要进行 RCT 就能通过某种方式发现降落伞可以大大降低跳伞运动中的死亡风险，而且吸烟和肺癌之间的联系最初也不是通过人为实验发现的。虽然我们可以从运行机制的背景知

识中把握事件发生的原因，但我们仍然要意识到，在某些情况下，实验可能也会导致我们得出错误的结论。比如以下两种情况：多个备选原因都能导致某种结果，或实验中的干预措施会带来某些副作用。

如果我们想知道某个基因会出现什么样的显性性状，通常会在测试中抑制这个基因的活跃性（基因敲除实验），然后观察这种显性性状是否还存在。这种做法的依据是，如果我们认为某个基因是导致某种显性性状的基因，并且在抑制了这个基因的活跃性之后，这个显性性状依然存在，那么这个基因就不是导致这种显性性状的基因。在这个例子中，我们假设导致某种结果的原因只有一个。然而，如果这个显性性状依然存在，那么可能还有一个备选原因也能导致这种性状的出现，在第一个基因被抑制后，这个备选原因会代替第一个基因起作用。很多生物学案例都是如此。为了保证某个性状的稳健性，可能会存在这样一个基因，它既能导致某种性状的出现，又能抑制另外一个基因；如果这个基因被抑制了，另外一个基因就会代替这个基因起作用。

如果原因被剔除之后，相应的结果不再出现，这也并不意味着我们已经找到了**真正的**原因。如果没有氧气，房子就不会失火，因为氧气是房子起火的必要条件。但我们不会因此认为氧气本身会引起火灾（它是不充分条件），因为火灾的发生还需要很多其他条件（比如热源和易燃物）。

假设我们想要证实长跑是否有助于减肥。为了测试这个观点，我们将试验的参与者随机分成两组，一组接受马拉松比赛的训练，另一组每周进行几次一两公里的长跑。矛盾的是，在针对这个假设的研究中，那些跑得多的参与者的体重不仅没有下降，反而还增加了。我们真正想要考察的是，在假设其他因素保持不变的情况下，长跑对体重有什么影响。但是，实验中的这种长跑导致了一些意想不到的后果。也许参与者在长跑之后觉得很疲惫，于是在不跑步的时间里，他们坐着的时间变长了。他们的饭量可能也增大了，从而超额补充了运动消耗掉的热量。因此，副作用不仅会

在我们试图使用原因来设计干预措施和政策时带来困扰，还会在一开始就阻碍我们找到正确的因果关系。更麻烦的是，因果之间存在两个本质上就不同的作用路径，而这两个路径可能会抵消对方的效果，或者会导致某种与预期的关系完全相反的关系。这正是我们在第 5 章讨论过的悖论，而且这种情况并不是观察性研究特有的现象。

所以，虽然实验法是寻找原因的一个好方法，但是我们并不一定要使用实验法来寻找原因，而且使用实验法也不一定能够找到原因。

注释

1. Grodstein 等（1997）。
2. Mosca 等（1997）。
3. Hulley 等（1998）。
4. 妇女健康倡议调查员组织写作组（2002）。
5. Woodward（2005）理论是提倡使用干预法来寻找因果关系的主要理论之一。
6. Holson（2009）。
7. 想要了解最近竞选活动中的 RCT 的更多内容，参见 Issenberg（2012）。
8. Green（2012）。
9. Lind（1757）。想要了解 RCT 在 Lind 之前和 Lind 之后的更多历史，参见 Bhatt （2010）。
10. Lind（1757），149。
11. 在一些 AIDS 药品试验中就出现了这种情况（Collins 和 Pinch，2008）。
12. 这一设计存在很多问题，其中包括如何保证各组之间的可比性，以及如何保证有足够数量的密集群体（Campbell 等，2004）。
13. Keeter 等（2008）。想要了解更多关于调查研究的内容，参见 Groves 等（2009）。
14. 对有些研究来说，IRB 也许不同意研究人员使用那些没有完成整个实验的参与者的数据，但也有一些指导性标准明确要求使用这样的数据以避免出现偏差。比如说，FDA 指导标准就要求在分析的过程中使用我们搜集到的那些没有完成实验的参与者在退出实验之前的数据（Gabriel 和 Mercado，2011）。
15. 想要了解更多失访问题，参见 Fewtrell 等（2008）。

16. 想要了解链霉素试验的历史，参见 Crofton（2006）。
17. 为了保证不同组别之间的性别平衡，人们在试验中用了几组信封来将男性参与者和女性参与者分开，并轮流打开来自每一组的对应性别的信封。
18. 想要了解更多这方面的伦理问题，参见 Macklin（1999）。想要了解关于病人想法的研究，参见 Frank 等（2008）。
19. 想要回顾关于安慰剂效应的研究，参见 Price 等（2008）。
20. 想要了解这方面的一些例子，参见 Kaptchuk 等（2010）。
21. Beecher（1955）。
22. Blackwell 等（1972）。
23. 想要了解关于这方面内容的简介，参见 Schulz 和 Grimes（2002）。
24. Noseworthy 等（1994）。
25. 三盲实验也可以指接受治疗的人、负责治疗的人和评估治疗效果的人都不知道实验小组是如何划分的。
26. Young 和 Karr（2011）。
27. 最近出现的新注册报表发布模型（Chambers 等，2014）就是这样一个例子。
28. 有一项研究将登记在案的抗抑郁药的临床试验和那些已经发表的临床试验进行了对比，发现研究结果和最终的发表行为之间存在高度相关性（Turner 等，2008）。
29. 它可能并不像看上去的那么牵强。参见 Boyd 等（2005）；Hajjar 等（2007）。
30. 想要了解这方面的例子，参见 Rothwell（2005）。
31. Heres 等（2006）。
32. 想要了解这方面的例子，参见 Moher 等（2001）。
33. Rothwell（2005）。
34. 想要回顾一下这方面的内容，参见 Kravitz 和 Duan（2014）。
35. March 等（1994）。
36. Kleinberg 和 Elhadad（2013）。
37. 想要了解在计算机科学背景下，人们关于这一差别的研究，参见 Drummond（2009）。
38. Prinz 等（2011）。
39. Begley 和 Ellis（2012）。
40. Young 和 Karr（2011）。
41. Klein 等（2014）。
42. Herndon 等（2014）。

43. 比如说，在人们发表了关于套利机会的学术论文后，有些这样的套利机会会变小（McLean 和 Pontiff，2015）。
44. 想要了解更多关于机械性因果关系的内容，参见 Glennan（2002）；Machamer 等（2000）。
45. Russo 和 Williamson（2007）。
46. 想要了解这方面的例子，参见 Charney 和 English（2012）；Fowler 和 Dawes（2008）。

第 8 章　解释

“这件事引起了那件事”
这句话意味着什么？

一名居住在堪萨斯州的男子在经历了一系列梦游事件之后，去了一家治疗睡眠障碍的诊所，想要查出他到底得了什么病。一个多月后，他被确诊为非快速眼动睡眠异常症。这种睡眠障碍可能会导致人们做出一些奇怪的行为，比如在睡眠中到处走动或吃东西等，但大脑不会记住这些事情。在他被确诊两个月之后，医生增加了他的用药量，而在增加用药量的两天之后，他被捕了，并且被控告杀死了自己的妻子。[1]

睡眠异常症患者意外杀人的案例十分罕见，但这个案例会是其中之一吗？有一些证据显示，这个案例可能真的是睡眠异常症患者意外杀人的案例。这名男子在被捕之前拨打了 911，他在电话里的表现非常奇怪，似乎对于已经发生的事情感到十分困惑。鉴于他有睡眠异常症病史，所以一切听起来就好像他还在睡梦中一样。然而，进一步调查之后发现，梦游时的暴力行为的很多常见特征在本案中并未出现。他和妻子有过争吵（梦游时的暴力行为通常没有任何动机），他们俩之间的距离很远（梦游者通常必须要靠近他人才会出现暴力行为），而且他使用了多种武器（梦游时的暴力行为通常只用一种武器）。最终，这个案子水落石出，被证实为一起谋杀事件。

这个案子的重点是，不能只因为睡眠异常症可能会导致谋杀，而这个案子里既有睡眠异常症也有谋杀，就理所当然地认为一定是睡眠异常症导致了这一起特定的杀人事件。

当我们询问某件事情为什么会发生时（比如为什么会发生某一场暴动，为什么两个人会发生车辆碰擦事故，以及为什么某个候选人会赢得选举），我们想要的是一个事件为什么会发生或者为什么未发生的因果关系解释。除此之外，还有一些其他类型的因果关系解释（比如解释两个事物之间的联系）和非因果关系解释（大部分都是数学方面的例子[2]），以及很多科学解释理论。在本章中，解释行为的目标就是要找到一些导致特定事件发生的原因（也就是实体原因，本章中的实体原因和因果关系解释是可以互换的同一事物）。大部分情况下，我们想要解释的似乎都是出了问题的事件，但我们也可以问一问人们为什么能够成功地避免某场核灾难，或者人们是如何成功让某种传染性疾病停止传播的。

类型层面上的因果关系让我们能够深入认识事物的一般属性，比如阳光照射会引起皮肤受伤；而实体层面的因果关系则与具体事件有关，比如马克 7 月 4 日没有涂防晒霜，然后在海滩上待了一整天，结果他的皮肤被晒伤了。在类型层面上，我们想要获得的是可以用来预测未来事件的知识，或者是可以用来在普遍意义上（比如针对整个人口群体的政策）改变事件发展进程的知识。而实体层面的因果关系则是关于某个具体事件的因果关系。比如我想知道为什么我的航班会晚点；而如果航班晚点其实是飞机机械故障造成的，那么仅知道天气和空中交通情况通常是导致航班晚点的原因，对我来说其实并没有多大的帮助。实体因果关系的重要性通常比它在这个案例中的重要性要大得多，例如，在划分法律责任的过程中，或者在根据各人贡献大小而颁奖的过程中，实体因果关系都起了非常重要的

作用。然而，有时可能会出现一些一次性事件，这种事件永远都不会发生第二次。在这种情况下，我们可能在事件发生之前都不知道还存在这样的因果关系。[3]比如法国和墨西哥之间的那场在某种程度上由甜品引发的战争，这种引发战争的原因闻所未闻。[4]药品的某些副作用或相互作用可能从来没有在临床试验中出现过，但当这种药品被用在更大且更加多样化的人群中时，可能就会出现这样一些副作用。

然而，这种特性恰恰导致了人们难以确定实体（也称为特定或实际）因果关系。如果我们不能把类型层面的原因当作实体原因，那么即使这些原因出现了，我们又如何才能得知某件事情为什么会发生呢?

本章要考察的是，在某个具体的场合，一件事引起了另一件事意味着什么？这样的因果关系和事物之间的普遍联系有何区别？在研究普遍联系时，我们寻找的是事物之间不受时间限制的属性。很多方法都能够帮助我们理解这两种类型的原因是如何组合在一起的。我们可以先找到事物的一般属性，然后将这些属性套用到具体事物上；也可以先从具体案例出发，然后得出一般性的结论；还可以提出与这两种方法完全不相关的研究方法。每一种方法都要求我们对已有的信息进行筛选和评估，但我们的研究领域一直在不断发展并试图实现这种解释的自动化。我们将会考察如何才能实现这种自动化，并且探讨实现过程中面临的一些挑战。除此之外，还将考察法律领域中的因果关系，探讨陪审员是如何根据证据进行推理的。法律案件面临着很多和其他案例一样的挑战。不仅如此，我们在法律案件中还必须做出裁决。陪审员们一方面要判断证据本身的可靠性，另一方面还要把那些分散的证据整合在一起，形成一个合理且连贯的案情分析。他们分析案件的这种方法可以指导我们分析其他案例。

8.1　寻找某个事件发生的原因

我们知道破旧的洗衣机会让水龙头漏水，但是仅知道这一点是否就能解释为什么上周二 Ann 家的水龙头会滴水呢？由于机场的安检队伍太长，结果 Bernie 没有赶上他乘坐的航班，我们是否可以由此推断出安检队伍是导致旅客误了航班的原因呢？在第一个案例中，我们使用了一般性的、类型层面的关系来解释某个具体案例。很多分析方法都是这样分析问题的。但是，我们也可以把很多具体的案例聚集在一起，然后总结出事物的一般属性。[5] 我们先使用类型层面的原因来解释实体原因，讨论一下这种分析方法面临的一些挑战，然后放松类型原因和实体原因之间的联系，最后，在后面几节中完全切断类型原因和实体原因之间的纽带。

8.1.1　出现多重原因时

假设我们想知道是什么导致了某一场车祸。虽然我们无法从一场车祸中找到某一条规律，但是可以使用我们的先验知识来解释这一场车祸。例如，我们可以使用 Mackie 的 INUS 条件（非必要充分条件中的非充分必要部分，详见第 5 章）来找到好几组导致车祸的因素，如果这几组因素中至少有一组因素的各个组成部分都出现了，那么车祸这个结果就一定会发生。但是，可能有多组因素都足以导致车祸这一结果，所以这几组因素中的每一组都不是必要条件。

如果我们想证实路面结冰是导致这起交通事故的实体原因，那我们还要知道令路面结冰导致交通事故的其他因素也存在于现场，比如能见度低。在这个例子中，路面结冰本身并不足以导致交通事故。但是，如果路面结冰和能见度低这两个因素都出现了，驾驶员又醉酒，而且交通也非常拥挤，情况又会怎样呢？根据图 5-2 所示，这些组合足以引起交通事故了。由于这个超定事件中出现了多重充分原因，如果使用 Mackie 的分

析方法，我们将无法找到事件发生的原因。

分析具体案例的另一种方法是假设法。假如路面没有结冰的话，这起交通事故还会发生吗？假如驾驶员没有喝酒的话，事情的结果会有什么不同吗？在这种分析方法中，我们将原因定义为某种能够改变事件发展进程的事物——如果这个原因没有出现的话，事情的结果将和我们知道的实际发生的结果大不相同。

这正是第 5 章讨论过的反事实推理法。前面说过，反事实依赖性是指：如果原因没有发生的话，结果也不会发生；如果原因发生了，结果也一定会发生。反事实推理法主要用于解释事件发生的原因，其核心思想是影响事件的发展过程。

反事实陈述随处可见：如果我没有吃这个药的话，我是不会康复的；如果我没有熬夜而是早点睡觉，那我就不会头疼了；如果我穿越街道的时候没有那么匆忙，我就不会被绊倒了。反事实推理的过程和我们解释某件事情为什么会发生（在心理学领域被称为归因）的过程很相似，[6] 但是反事实推理并不能完全解释归因过程。在有些情况下，反事实推理法认为事件之间不存在因果关系（但是人们并不赞同这样的结论）；但在另一些情况下，虽然人们认为两个事物之间不存在因果依赖性，但反事实推理法却发现它们之间存在反事实依赖性。

有一项研究测试了这两种推理之间的联系。在这项研究中，参与者读了一个故事，故事中的主人公被人下了慢性毒药，但在毒药发作之前，他在过马路的时候遭遇了车祸。[7]故事中说，这个人一辈子干的坏事太多，所以才会遭遇这些杀身之祸。读完故事后，研究人员让参与者判断故事中的主人公死亡的原因是什么。这个故事中的两个原因（毒药和交通事故）都可能导致死亡，所以我们无法通过反事实依赖性来判断他的死亡原因。但是，参与者并不认为这些原因是对称的，事实上，他们认为交通事故与主人公的死亡更加相关。然而，当研究人员让参与者使用反事实推理法或

归因法进行原因分析时，他们给出的答案却截然不同。由此可见，这两种推理过程是不一样的。尽管参与者并不认为主人公犯下的罪行是他死亡的原因，但他们认为从反事实推理的角度来看，这是导致他死亡的最重要的因素。他们可能认为，如果时光倒流，这个因素可以改变的话，事情的结果会大不相同。[8]

但是，也不是所有人都是这样想的。上面说的是最普遍的结论，并不是所有的参与者给出的答案都是一样的，要记住这一点。这些结论是通过因果判断或反事实推理所得出的最常见的结论，但还有一些参与者得出了完全不同的结论。我们随后将会讨论陪审员们在审判案件的时候是如何进行推理的。在这样的推理过程中，我们关注的核心问题是：为什么人们在分析同样的事实时会得出不同的因果结论？我们想知道人们是如何思考的，以及哲学理论和人类判断之间出现分歧的原因是什么。但是，我们并不清楚在人类判断出现分歧的时候，是否还能利用哲学方法来获得同样的认知。第 2 章和第 3 章讲过，我们寻找和评估证据的方法都是有偏差的，而且不同的人可能会有不同的偏差。

在某些案例中，我们可以说多重因素共同导致了某个结果，但是在另一些案例中，我们却不得不进行责任划分。对于一个行刑队而言，可能所有开枪的队员都是导致犯人死亡的原因，但我们不需要知道致命的一枪究竟是哪个队员开的。但在法律案件中，我们需要根据每一个因素对原告造成的伤害程度来划分其应该承担的责任比例。假设一个人由于长期在噪声很大的环境中工作并且脑部受了外伤，从而丧失了听力，而另一个人完全是由于工作场所的噪声而丧失了听力，那么这两种情况下法院所判的赔偿是不一样的。而且，法院判给第一个人的赔偿还需要由这两个原因的责任方按照责任比例共同承担。但是在现实生活中，我们无法确切地知道

（比如说）这个人丧失的 40%的听力是工作场所的噪声导致的，另外 60%是脑部外伤导致的。

有人建议，当我们无法确定某个因素是否应该承担责任时，可以这样来划分责任比例：根据每个因素在整个人群中导致的某种结果的比例来确定这个因素所应承担的责任比例，或者根据这个因素相对于所有潜在的风险因素而言能够导致某种结果的比例来确定这个因素所应承担的责任比例。[9] 但是，这种建议假设了事件发生的一般性概率可以直接适用于某一个具体事件，但事实上，我们无法确定这种比例是否对每一个人都是不同的。我们在使用一些方法来计算具体案例的发生概率方面已经取得了一些成绩，但是这些方法要求我们对事件发生的背景知识有着充分的了解。

更加具体地界定我们想要解释的对象，能够解决表面上的超定案例。比如说，在目前所讲的所有案例中，我们一直都把各种死亡事件看成是同一类型的事件。我们并没有区分下午两点的交通事故中的死亡事件和晚上十点中毒导致的死亡事件。每个人最终都会死，所以我们认为死亡早晚会发生，但是有些事情却导致它发生的时间提前了。

因此，在运用反事实推理法时，不要只看结果会不会发生，而是要看结果是否会以不同的方式发生。假如这个案例中的受害人没有遭遇车祸或者体内的毒药没有发作，那么他也有可能会在不同的时间、以不同的方式死去。[10] 通过这种方式，我们能够发现一些在其他情况下的表面上的超定事件发生的原因。

8.1.2 解释可能具有主观性

如果我们想要知道前面那个例子中的受害者为什么会死，可能会想知道这几个问题：为什么死的偏偏是这个受害者而不是那个犯罪分子？为什么这起交通事故会致死？为什么受害者偏偏死在这一天而不是那一天。

也就是说，即使我们解决了超定问题，还必须考虑两个人使用同一

推理方法可能也会得出不同的因果结论。我们选择的测量对象和描述测量对象的方式（比如体重与身体质量指数）都会影响到类型层面的推理。同理，这些选择可能也会影响到实体层面的解释。除了这些针对变量的选择以外，我们还必须确定哪些因素出现了、哪些因素没有出现，这无疑增加了实体层面推理活动的复杂性。

有人可能认为酒驾是个非真即假的变量，并且很多数据都能证明这个变量是否为真。但是，有人一年只会去听一次非常喧闹的演唱会，而有人则是摇滚乐队的成员或者每周都会去听一次非常喧闹的演唱会，那么对于这两种人而言，由于噪声而丧失听力的风险是不一样的。同理，驾驶员醉酒的程度也是不一样的。这种程度差异对于解释行为和因果推理的影响是有差别的，在因果推理过程中，我们是从数据中来界定一组变量（比如将身高和体重转换成身体质量指数），然后从这组变量中寻找变量之间的关系。

在实体因果关系案例中，我们将实际情况与我们掌握的类型层面的知识联系在了一起。可能之前有一项研究发现运动量大的人静态心跳率比较低，[11] 但现在我们想知道的是，Tracy 的静态心跳率比较低是否是运动量大导致的。如果幸运的话，先前的研究可能会准确地告诉我们一个人必须锻炼多少次、每次锻炼多长时间（比如一周 6 次，每次 30 分钟）才能降低静态心跳率。但除此之外，我们还需要考虑很多问题。只有锻炼三个月以上的人身上才会出现这种关系吗？所有的锻炼形式都一样吗？要不要单独考虑瑜伽和游泳这两种锻炼形式？如果 Tracy 只在天气暖和的时候才锻炼，在冬天的时候根本不锻炼，这会影响这种关系吗？我们之所以要将实体层面的观察和类型层面的知识结合起来，是因为在确定事实真相的过程中，人们可能会不自觉地带入自己的主观性。[12]

不同的人对于同一件事可能会提出不同的问题，而且他们认为的重要因素可能也各不相同（可能是因为他们所能控制的因素各不相同），但

这并不会改变每一个因素在整个事件中真正起到的作用。比如说，某个人赢得了诺贝尔奖是因为这些因素的共同作用：勤奋、幸运、早期在学校接受的自然科学教育，可能还有前面提到过的巧克力。如果有人专门去研究诺贝尔奖和巧克力之间的联系，那这只会改变研究者可能会问的问题，而不会改变巧克力对人们获得诺贝尔奖的贡献是否要比运气的贡献更大这种事实。不过，当我们想让解释自动化的时候，就必须减少主观判断，而且要找到那些最重要的因素。要想解释长期在巨大噪声中生活对人们的影响，我们就要去了解某个人在噪声中生活的经历，所以，我们需要掌握的数据可能有这些：这个人每周听演唱会的次数、这个人的工作场所是否有噪声，或者这个人是否生活在建筑工地附近。

8.1.3 原因出现的时间

如果我们假设这场交通事故是酒驾引起的，那么在事故发生时，驾驶员应该处于醉酒状态。而对于那些潜伏期很长的传染病来说，我们则会假设病人一定是在过去某个时间接触了传染病毒。不过，某个人的流感不可能是从一年前和他一起吃过一次午餐的流感病人身上感染的，也不可能是从一分钟前和他一起吃过午餐的流感病人身上感染的。

因此，使用类型层面的原因来解释实际案例时，让问题复杂化的第三个因素就是时间。即使我们掌握的类型层面的信息并没有告诉我们某个原因需要多长时间才能导致某种结果，我们依然不可避免地要考虑到时间因素，因为时间因素决定着哪些信息与实际案例有关。如果我们对因果关系中的时间因素一无所知，就不得不做一些判断来决定某件事情是真还是假。比如说，如果我们想知道某个人得流感是否是因为他接触了流感病毒，那么这个人接触流感病毒的时间就很重要，它会告诉我们这个人这次的流感是否可能是那次接触的流感病毒引起的。

有些因果推理方法会在原因和结果之间留一个时间间隔或时间窗，

让我们能够认识到患小儿麻痹症可能会导致病人在康复十五年后出现小儿麻痹后遗综合征。[13]知道了这一点，如果病人在刚刚康复了几个月之后就出现了一些小儿麻痹后遗综合征的症状，那么这些症状是否是由小儿麻痹后遗综合征引起的就毋庸置疑了。如果病人是在我们知道的时间间隔之内患了小儿麻痹症，那么我们就可以用所了解的类型层面上的因果关系来解释这个实际案例了。而且，如果两个人使用的数据相同，那么他们对于"小儿麻痹症是否可以解释病人身上出现的症状"这个问题应该会有相同的观点。

时间对因果关系的影响还不止于此。假设有一种药物可以在 30 分钟到 60 分钟内减轻头痛症状。Charlie 得了头痛症之后吃了这种药物，结果 62 分钟之后他的头痛症状减轻了。那么，这种药物对 Charlie 的头痛症有没有帮助呢？尽管 62 分钟不在 30 分钟到 60 分钟的时间窗内，但如果由于症状消失的时间与我们了解的时间窗不完全一致，就说这个药物不可能是 Charlie 头痛症状减轻的原因，那我们对时间的要求似乎过于苛刻了。以我们对头痛药物的了解，加上我们服用头痛药物的经验，药物起作用的时间窗不可能刚好只有 30 分钟，不可能在第 29 分钟的时候还没有起任何作用，然后到了第 30 分钟就立刻起作用了。时间窗可能是某个原因起作用的主要时间段，所以它并不一定意味着某个结果**不可能**在时间窗以外的时间发生，只是说这个结果在时间窗以外的时间发生的可能性很低而已。与之相反的情况是登革热，这种传染病可能会突然暴发。通过研究登革热的历史数据，我们可以得知已经观察到的这种疾病的最短和最长潜伏期是多久。在这种情况下，我们就能更有把握地确定某个人不可能是因为在上述潜伏期以外的时间接触了登革热病毒而感染了登革热。

尽管 Charlie 的案例与我们先前了解的知识并不吻合，但这种情况与我们的先验知识非常接近，所以我们很想让我们评估解释的方法变得灵活一些，以便可以将 Charlie 服用的药物认定为他头痛消失的原因。同时，我

们还要有能力处理那些时间要求更为严格的案例。因此，在找到那些类型层面的因果关系时，要能够清楚地表明这些时间窗到底是某个结果可能会出现的唯一时间段，还是某个结果最有可能出现的时间段。时间上的灵活性还反映了这样一个事实：我们所知道的时间段只是根据某些先前的数据或先验知识而得出的结论。假如我们的结论来自于一个小的数据集，那么我们可能观察不到某个很不常见的、极为短暂的潜伏期。或者数据测量点之间的间隔很大，甚至导致第一次跟踪研究要到两天之后才能进行。在这种情况下，由于数据粒度问题，我们可能永远也不会知道这个疾病有没有可能在第一天就发作。

此外，如果我们所了解的事件在实体层面上发生的时间可能是错误的，那么严格地遵从某个已知的时间窗就没有任何意义了。如果我说某件事情发生在一个星期前，那么我说的一个星期前既可能是指 6 天前，也可能是指 7 天前或者 8 天前。同理，“1 年前”几乎不可能是指“正好 365 天前”。所以，即使我们知道某件事会在一年内导致另一件事，严格遵循时间窗的限制也会导致我们忽略数据内部的不确定性。[14]

8.2 具有不确定性的解释

解决上述问题的一个办法就是放松类型层面的关系与实体层面的关系之间的联系。出于很多原因，我们观察到的东西和我们已有的认知并不吻合。因此，我们可以将这种不确定性融入我们的解释过程当中。有人在服药 29 分钟之后头痛症状就消失了，有人在服药 290 分钟之后头痛症状才消失，与第二个案例相比，第一个案例的药物更有可能是头痛症状消失的原因。我们有时可能对实际发生的事情不是很有把握，这时也可以利用这种不确定性来进行更加准确的解释。也许我们并不确定 Charlie 有没有服用扑热息痛（一种解热镇痛药），但我们看到有一盒打开的扑热息痛放

在一杯水旁边，于是就可以使用这些间接信息来估算他服用药物的概率。在此不详述这种方法，但这种方法的基本原理直接代表了我们的先验知识中的不确定性和我们对实体案例中的信息的不确定性。[15]

Mackie 的 INUS 方法假设了我们对事物的运行机制有着充分的了解，所以能够界定确定性因果关系复合体，比如一组因素的出现总能带来某个结果。但是，很多因果关系的出现都是有一定概率的（这可能是事物本身的不确定性导致的，也可能是我们对世界的认知不够全面导致的）。即使某个原因导致某种结果的概率很低，但它在实体案例中仍有可能成为导致某个事件发生的原因之一，而我们计算出的概率或原因强度可以告诉我们这种情况发生的可能性有多大。然后，我们可以利用这些砝码来评估各种因果解释的依据。[16]

我们来看看这种方法是如何运作的。假设我们想知道 Irene 昨晚为什么会失眠。我们测量了各个原因的显著性值（详见第 6 章），然后发现喝 4 盎司的浓缩咖啡在 4 小时内导致失眠的显著性值为 0.9。如果我们了解到 Irene 睡前 3 小时曾去过一家咖啡店，并且喝了 4 盎司浓缩咖啡，那么喝浓缩咖啡导致她失眠的显著性值就会是 0.9。但是，如果她从咖啡店回来之后又熬夜看了一会儿电视，实际上是在喝咖啡 6 小时之后才出现失眠的，那么由于失眠发生在浓缩咖啡影响睡眠的时间范围之外，因此浓缩咖啡导致失眠的显著性值应该比 0.9 要低一些。图 8-1 展示的是这些事件发生的序列以及这一因果关系的已知时间窗（灰色部分）。6 小时位于灰色长条所示的已知时间窗之外，所以 Irene 那时候的失眠似乎不可能是之前喝浓缩咖啡导致的。

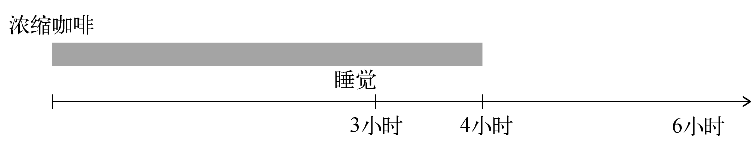


图 8-1　喝浓缩咖啡导致 4 小时内失眠

从直觉上讲，我们并不认为人们在喝了浓缩咖啡后 0 小时到 4 小时内失眠的概率是一成不变的，或者一过 4 小时失眠的概率就会骤降为 0。相反，实际情况可能更像图 8-2 所示的那样，在过了第 4 个小时后，浓缩咖啡导致失眠的概率就会慢慢降低。当我们对同一个原因在不同的时间段对于结果的显著性值进行加权时（或者在解释某个原因对不同时间段的影响时），应该将这种概率和显著性值结合在一起考虑。这意味着，一个对结果影响比较大的原因即使和已知的时间段不太吻合，它的显著性依然比一个对结果影响比较小但是实际出现的时间段与已知时间段完全吻合的原因更大。如果 Irene 睡觉时房间里有点太暖和了，就可能会增加她睡不好觉的概率。但是，我们可能依然会认为 4.5 小时前喝浓缩咖啡才是导致她失眠的罪魁祸首。

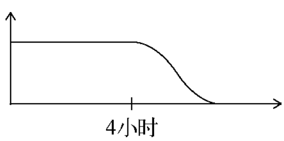


图 8-2　失眠的概率随着时间的变化而产生的变化。横轴表示的是喝浓缩咖啡后的小时数

这种方法的基本思路是，根据我们所掌握的实体层面的信息来对类型层面的显著性值进行加权处理。对于一个具体案例而言，由于各个事件发生的时间不同或者具有不确定性，某个因素的显著性值要低于它在类型层面的显著性值。我们可以根据已知的事物运行机制（比如某种药物的作用机理）或者先前的数据（只需计算随着时间的变化某个结果出现的概率）来定义一个函数，让这个函数来告诉我们如何将观察到的数据与某个原因仍在起作用的概率相匹配。图 8-3 展示的是这个函数的几个例子。在图 8-3a 中，概率的值只有两种可能：0 或 1。这意味着某个原因只有在时间窗显示的这个时间段内才可能会导致某种结果；在时间窗以外的时间段，这个

原因不会产生任何显著性影响。相反，在图 8-3c 中，在时间窗以外的时间段内，某个原因导致某种结果的概率下降的速度要慢得多。这种方法不再主观地判断某个具体案例是否符合我们对类型层面的认知，而是将类型层面的因果关系和实体层面的因果关系更有条理地结合在了一起。

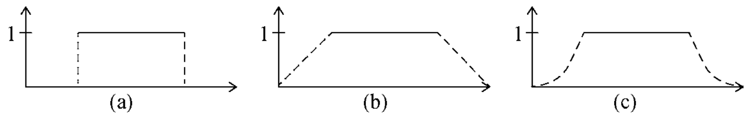
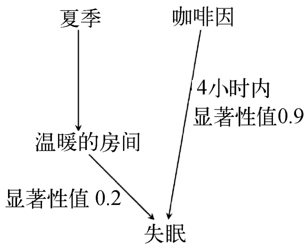


图 8-3　各种可能出现的对观察到的时间段和已知的时间段进行加权的函数。实线表示某个原因最有可能导致某个结果的时间段，虚线表示在已知时间段前后，某个原因导致某种结果的概率是如何变化的

如果我们并不确定 Irene 有没有喝浓缩咖啡，又会怎么样？我们可能得知的信息有：她在咖啡店见了一个朋友，她通常会喝很多浓缩咖啡，但有时却只喝不含咖啡因的茶。在没有直接知道某个原因是否出现的情况下，我们可以使用其他信息来计算这个原因出现的概率，然后再次对类型层面的信息进行加权处理。所以，如果我们可以肯定某个原因已经发生了，那么这个原因在实体层面的显著性值就等于它在类型层面的显著性值；相反，如果根据我们掌握的观察数据，某个实体层面的原因发生的可能性不大，那么这个原因的显著性值也会相应降低。

在这种情况下，我们看到的是一组原因和一系列事件，并且要将这两方面的信息结合起来，从而确定各种假设的显著性值。[17]我们由此得出的结论不再类似于“这个原因导致了（或没有导致）那个结果”这种二元性结论，而是对各种可能的原因进行的排序，如图 8-4 所示。一个结果会有很多可能的因果解释，我们在测量每一个解释的显著性值时，都会将类型层面上的因果显著性值、时间段的吻合程度以及每一个原因在各个时间段发生的概率结合在一起考虑。与其他方法不同的是，这种方法不需要完

全了解哪些变量是真的、哪些变量是假的，而且实体层面的时间段也无须与类型层面的时间段完全一致，这让我们能够更好地处理像因果关系链和超定这样的案例。



（a）类型层面的因果关系

事件：	时间：
咖啡因	下午4点
夏季	一整天
温暖的房间	未知
失眠	晚上10点

（b）实体层面的信息

原因：	显著性：
咖啡因	0.75
温暖的房间	0.15

（c）显著性值

图 8-4 结合类型层面的关系和实体层面的信息来解释失眠这个结果，对各种原因进行了排序

8.3 将类型层面和实体层面分开来看

假设我们找到了一组导致篮球进篮的因素。某个周六下午，一名篮球运动员投球时，这些进篮的因素都出现了，但在最后一分钟却因为一场地震而未能投进。虽然所有导致篮球应该进篮的因素都出现了，但篮球却没有进篮。这些因素没有让篮球进篮，但是，除了地震这个因素以外，其他因素也没有导致篮球不进篮。

到目前为止，我们主要是在解释实际发生的事情为什么会发生。我们在第 2 章分析心理学文献时曾经讨论过一件非常奇怪的事：人们可能会因为一些没有发生的事情而受到责备。有人可能会犯下谋杀未遂罪，即使某人考试作弊未遂也仍然应该受到谴责。但是，如果某人没有替你浇花，但花依然活着，我们又该如何解释这种现象呢？这个花本来应该已经死了，但是它却没有死。发生了花缺水的事件，但是它却不是导致花活下来的原因。

第一天没有给花浇水，花幸存下来的概率降低了。随着不给花浇水的时间越长，花幸存下来的一直直线降低。虽然某个因素的出现让一件事发生的可能性降低了，但是这件事还是发生了。那么，从直觉上来讲，这件事的发生并不是某个因素导致的，而是在出现了某个不利因素的情况下仍然发生了。同理，虽然某个因素的出现让某件事发生的可能性提高了，但是这件事仍然没有发生，那么，这件事之所以没有发生也不是某个因素导致的，而是在出现了某个有利因素的情况下仍然没有发生。比如说，尽管我们拥有良好的医疗服务条件，但是某个病人仍然有可能死亡。

改变某个结果出现概率的原因有很多，但在这个结果实际出现的时候，并不是每一个原因都对这个结果的出现产生了影响。在某些情况下，一个事件可能会提高某个结果出现的概率，却不会导致这个结果的出现。比如说，假设 Adam 和 Betty 都得了流感。他们俩在相距一周的时间里分别和 Claire 一起吃过午饭。Claire 在第二次和他们其中一人共进午餐后的第二天就得了流感。Claire 和 Adam 一起吃过午饭后，她得流感的概率增加了，但随着潜伏期的延长，她得流感的概率又降低了；她和 Betty 一起吃过饭后，得流感的概率又提高了，而且一直到她真的得了流感为止，她得流感的概率一直很高。（如图 8-5 所示）虽然这两个事件都是类型层面的原因（与流感病人的接触），但这却不是一个超定的案例。相反，只有其中的某一次接触是导致流感的原因。在上一节中，我们使用了类型层面的时间段来处理这样的案例，但这一节所用的方法与之不同。这一节研究的是实体层面的概率是如何随着时间的变化而变化的。这种方法还能够处理一些实体层面的概率不同于类型层面的一般概率的案例。

通常情况下，疫苗是可以预防死亡的，但在极少的一些情况下，疫苗却是死亡的原因。虽然从来没有任何植物因为被浇了咖啡而死掉，却可能会有某种植物因为被浇了咖啡而死掉。即使受害者在一次谋杀事件中幸免于难，但我们仍然可以对谋杀未遂的凶手追究一定的责任。到目前为

止，我们考察的所有事件都有一个很关键的局限性，就是我们一直在依赖一般性的信息来解释具体的案例，并且假设某种因果关系在类型层面的显著性与实体层面的显著性是一致的。

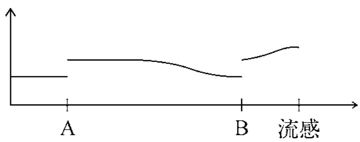


图 8-5　感染流感的概率随着时间的变化而变化。第一次与流感病人共进午餐后，感染流感的概率上升了，第二次与流感病人共进午餐前，感染流感的概率下降了。在第二次接触流感病人后，感染流感的概率一直上升到真的感染了流感为止

哲学家 Ellery Eells 提出了一个研究概率变化的方法：观察在某个原因出现之后，某个事件发生的概率是如何变化的，并且这个概率是如何随着时间的变化而改变的。[18]我们并不打算在此详细讨论这种研究方法，只简要介绍一下其主要特征：研究具体案例发生概率的方法要与研究一般性案例的方法有所不同，这种方法研究的是事件实际发生的概率是如何随着时间的变化而改变的。通过研究我们想要解释的具体案例发生的概率，我们能够将一般会发生的事情和实际发生的事情区分开来，并且意识到一个一般情况下可以预防某种结果出现的原因也可能会成为导致这种结果出现的原因。

最重要的是，这种方法能够修正我们的分析，让分析结果能够符合观察到的内容。Eells 曾举过这样一个例子：淘气的松鼠一般会将高尔夫球踢到远离球洞的地方，但在某一个案例中，有一只松鼠却将球对准球洞踢了进去，从而帮助了那名高尔夫球员。如果我们使用的方法是基于类型层面的概率，那么即使我们实际上看到球的运行轨迹让球落入球洞的概率越来越大，然后又看到这个轨迹在被松鼠踢了之后是如何发生变化的，我

们也无法修正先前所了解的类型层面的知识来将这种新的情况考虑进去。结果，我们就会得出脱离实际的、与直觉相矛盾的结论。

在一个事件发生后，另一个事件发生的概率开始上升，并且一直上升到真正发生为止，那么人们就会认为另一个事件的发生是由第一个事件导致的。相反，如果一个事件发生后，另一个事件发生的概率下降了，而在这种情况下另一个事件还是发生了，那么人们就会认为这个事件是在尽管有不利因素的情况下仍然发生了。[19]但是，由于我们很难得知一些信息，比如某个高尔夫球在其运行轨迹的每一个点上落入球洞的概率，所以在实际运用这种方法时会遇到一些困难。

8.4　使解释过程自动化

我们如何才能验证与事实相反的情况？如何才能得知某件事情发生的概率是如何随着时间的变化而变化的？很多备受推崇的哲学理论都有一个局限性，就是那些真正能反映类型层面与实体层面差异的理论往往要求我们对研究的情形有足够的了解，而且这种要求有时是不切实际的。如果我们能知道在某个时刻高尔夫球落入球洞的概率为 0.5，而在球被松鼠踢了之后，其落入球洞的概率增加到了 0.7，那这个信息对我们就很有帮助。但问题是，我们什么时候才能获得这样的信息呢？

如果我们能够为研究的系统建一个模型，那就可以解决上述问题了。根据一些简单的物理学知识以及我们对风和其他影响因素出现的可能性的一些假设，我们可以预测高尔夫球在被踢之前和被踢之后的运行轨迹。由于结果通常是不确定的，所以我们可以对球的每一个位置进行多次模拟，从而计算出球从那一点出发之后落入球洞的概率。当球离球洞很远时，风或其他不太可能出现的事件导致球的运行轨迹发生变化的概率很高，但当球靠近球洞时，要想让球偏离球洞，就必须出现更大的变故才行。使用

反事实推理法，我们可以模拟其他可能出现的情形，从而从数量上来测量各种情形之间的相似程度，以及在某个原因没有出现的情况下，某个结果出现的概率。

在医学领域，一般情况下，我们并没有足够的信息来如实模拟各种疾病有可能出现的发展过程。但是，我们可以使用来自其他病人的时间序列数据。假设我们想知道，一个病人在被确诊为肺炎的两周后存活了下来是否是因为服用了抗生素（也就是说，我们想要确定抗生素能否解释病人存活下来的事实）。那么，在服用抗生素之前，我们要利用我们所了解的关于这个病人的所有数据来搜集与这个病人具有相似病史的病人的信息，并计算出那些病人在两周后存活下来的概率。然后，只要将其与一开始就接受了抗生素治疗的那组病人的存活率相比较，就能看出病人在服用了抗生素后的存活率发生了什么样的变化。在前面的案例中，我们限制了高尔夫球的运行轨迹（一旦高尔夫球到达某一个位置，我们就只考虑它从那个位置出发后的运行轨迹）；在这个案例中，随着时间的变化，我们将以同样的方式来限制用于对比的病人群体。

从数据中寻找类型层面的原因一直是计算机科学研究中的一个主要领域，但关于解释过程自动化的方法的研究却没那么受关注。[20] 与自动化解决方案相比，人们更愿意使用因果推理来解释事物之间的关系。之所以会出现这样的情况，有一部分原因在于，我们很难将反事实推理这样的方法转换成计算机可以执行的指令。要想设计出一个程序，让它能够接收一些关于某个情形的信息，并且告诉我们导致某个结果的原因是什么，那就需要将解释原因的过程通过编码转换成一系列无须人为判断或主观想法就可以执行的步骤。第二个关键问题是：要如何去评估这些程序？想要知道一个计算程序是否有用，我们需要将它得出的结果和正确答案做对比。然而，对于实体因果关系来说，我们并不总是能够知道正确答案是什么。如果有一种方法可以用来确定不同的因素在某个结果出现的过程中扮演

了什么样的角色（比如确定两个独立的危险因素在导致一个人生病的过程中所应承担的责任的比例），而我们想要评估这种方法，这时我们是没有正确答案作为参照的，这一点特别让人头疼。

8.5　法律活动中的因果关系

本书开头介绍了一个案例，在这个案例中，由于人们误用了概率，未能理解因果关系的本质，结果导致 Sally Clark 蒙冤入狱。但是，除了虚假的统计数据以外，上诉法庭又是如何做出不同判定的？为何陪审员们在听到同样的证据后，商议了好几个星期也无法达成一致意见？

理解法律活动中的因果关系，[21] 特别是理解陪审团是如何做出判定的，这有助于我们更好地评估其他情境中的证据。在法律活动中，人们需要处理的是大量十分复杂而又相互矛盾的信息、事件的整个发展过程而不只是一个原因和一个结果，以及紧密相连的信息（人证如果说了一句错误的证词，有可能会降低他其余证词的可信度）。

有些哲学理论认为某些案例是无法解决的，比如超定事件。但在法律活动中，我们不能接受这种说法，因为我们无论如何都要做出判定。如果一个人既接触了石棉，又吸入了香烟的烟雾，那就要确定这两个因素中在这个人得肺病的过程中分别应该承担的责任的比例。如果这个人将获得赔偿，那我们必须划分出这些过错方所应承担的责任的比例。

在医学或历史学领域，专家们从多年的培训或经验中获得了一些技能，他们运用这些技能来解释病人身上出现的不寻常症状，或者找出某场政治运动为什么会在某个特定时间发生的原因。与这些专家不同，陪审员并不是法律方面的专家，也不是他们听审的案件细节的专家。正是由于这一点，法律活动中的因果推理才格外让人感兴趣。陪审员们可能不得不去评估环境因素和医学上的证据，以便确定癌症疫情密集暴发是否是一件不

寻常的事。虽然他们不是肿瘤学家或遗传学家，但他们却不得不去确定DNA证据能否明确指出导致癌症疫情密集暴发的嫌疑人。因此，陪审员们的推理活动和我们日常生活中的推理活动更为相似。在日常生活中，出于各种实际的原因，我们常常需要解释一些问题，但又不一定非要对相应的领域有十分深刻的了解。

8.5.1 要不是因为……

假设一名司机未能及时踩刹车，结果撞上了另一辆车。但司机不知道汽车的刹车实际上早就失灵了，所以即便他踩了刹车，也无法把车停下来。这个经常被引用的案例来自于一个真实的法律案件。在那个案件中，汽车租赁公司未能合理保养并检查汽车的刹车。[22]

人们之所以经常使用这个案例，是因为在法律案件中用来确定因果关系的核心方法之一是建立在反事实推理的基础之上的。在法律案件中，我们会问“要不是某个人的行为（或不作为），这一结果会出现吗”，比如说“要不是电工让电压激增，我的硬盘就不会受损”。这种推理也被称为“事实因果关系”，与反事实推理法完全一样。这种方法假设原因制造了差异，没有原因结果就不可能出现。然而，反事实推理法中的所有问题也会出现在“要不是”这一推理方法中。就法律案例而言，使用这种方法的主要障碍是无法处理超定问题。如果那个电工是在中午的时候胡乱改动了电压，但我的浪涌电压保护器也坏了，那么即便电工没有乱动电压，我的浪涌电压保护器可能也会损坏我的硬盘。硬盘损坏这一结果可能无论如何都会出现，既有可能是电工造成的，也有可能是浪涌电压保护器损坏造成的。所以，要是使用“要不是”推理方法，这两者都过不了关。

再回到交通事故的案件中，这起交通事故是由两个缺位（没有踩刹车以及没有确保刹车的可靠性）超定的，这两个缺位中的任何一个都会导致事故的发生。由于司机未曾试图踩刹车，所以尽管刹车有问题，但

是它并没有机会导致事故的发生。如果司机踩了刹车，那这件事就不是什么大事，但因为司机不知道刹车有问题，而且也没有采取恰当的安全保护措施，所以这个案件中应该承担责任的是司机。[23]

在超定案例中，两个或两个以上的因素都可能是导致某个结果的原因，其中任何一个都无法被判定为导致某种结果的唯一原因。相反，在优先权案例中，可能导致某种结果的因素有两个，但实际上起作用的因素只有一个。比如说，一个病人得了致命的疾病，在他死于疾病之前，护士先撤去了各种帮他维持生命的设备。

有一项针对 30 名法学院新生的入学调查。调查的问题是：在那起刹车有问题的交通事故中，过错方是谁？占比最高（43%）的回答是刹车和司机应该共同对这起事故负责。还有 33%的同学认为过错方在司机，23%的同学认为过错方在刹车。[24]有些给陪审团的指示中也明确提出了这个问题：在这种超定案例中，两个因素都可以被看成是导致某种结果的原因；或者陪审员应该更加详细地考察原因造成的结果，就像 Lewis 修改后的方法。Lewis 提出，如果两名纵火犯分别放了两把火，那这两把火吞没房子的速度比只有一把火要快得多，由此导致的结果不只是房子被烧毁那么简单，而是房子在 30 分钟内而不是在 90 分钟内被烧毁的问题，而这种时间上的缩短可能就是人们无法将火扑灭的原因。[25]

在上述案例中，常规的反事实推理法是没有用的。如果使用常规的反事实推理法，我们会发现这两名纵火犯都不是房子被烧毁的原因（因为总有一个备用原因）。然而从直觉上来看，两名纵火犯似乎都要承担一些责任。反事实推理法的缺陷之一在于，这种方法是将每一个原因分开考虑，而不是将每个原因当作导致某种结果的整体背景中的一部分。因此，Richard Wright 提出了一个叫作 NESS（充分条件组合中的必要成分）的理论框架，这个框架与 Mackie 的 INUS 条件类似，[26]其主要思想是：如果某个事件是一个充分（sufficient）条件组合（set）中的必要（necessary）成

分（element），那么这个事件就是一个原因。这意味着，如果整个充分条件组合都出现了，那么结果就一定会出现，而原因只是充分条件组合中的一个成分；相反，如果充分条件组合中缺少某一个成分，那结果就一定不会出现。在汽车交通事故案例中，刹车失灵是整个充分条件组合中的一部分，而没有踩刹车则是充分条件组合中的另一部分。那么，这两者都是 NESS 条件。根据 NESS 理论框架，它们似乎都难辞其咎。

然而，在这个案例中，要想找到正确的答案还需要考虑因果推理以外的因素。当我们说“考虑到驾驶员当时所掌握的交通知识，他应该按照某种特定的方式来采取相应的行为（即使他采取的行为并不会改变撞车的结果）”，我们其实是将这起交通事故的过错归到了驾驶员身上，因为他没有按照道路交通规则采取相应的措施。这就又回到了我们在第 2 章讨论的责任划分的问题上了——人们似乎会考虑当事人有没有违反某个行为规范。

8.5.2　近因

假设有人吓走了一只鸽子。鸽子飞走的时候，惊到了一位正在穿越街道的路人。路人驻足观望，结果导致一辆正在朝他骑过来的自行车不得不在最后一秒急转车头。自行车避让行人后，正好骑到了一辆出租车行驶的车道上。出租车为了避让自行车，结果撞上了一个消防栓。消防栓出水导致附近一栋大楼的地下室被淹，破坏了地下室的供电设施。虽然吓走鸽子是启动整个原因链的原因，我们也可以认为是吓走鸽子这件事导致了后面的一系列事件，但很少有人会认为吓走鸽子的那个人应该对之后出现的一系列事件负责——即使很多人都同意是那个人引起了这一系列的事件。因此，一个没有责任方的事故仍有可能存在一个原因。

除了要考虑“要不是”原因并进行 NESS 测试以外，我们还需要掌握原因和结果之间的距离，以便解释原因和结果之间发生的那些有可能干预

并改变结果的中间事件。近因就是和结果直接相连的原因。法律上的近因还具有可预见性，也就是说，人们应该能够预见该原因可能会导致某个结果。但在吓走鸽子的案例中，情况并非如此。所以，吓走鸽子这件事可能是一个“要不是”原因，却不是近因。

近因才是我们用来区分因果关系和责任的因素。[27]将责任局限于近因可以防止人们将其归咎于那些遥远的事件。那些遥远的事件可能触发了一系列事件，但是最终导致的结果却是无法预见的。前面说过，可传递性是反事实推理等方法中的一个主要缺陷。我们除了会发现遥远的原因以外，还可能会发现某件事情有时能够避免某个结果，但实际上又通过另一种方式使其发生了。比如下面这种情况：由于出租车司机车开得太慢，导致你错过了一场晚宴，那场晚宴中的所有人都出现了食物中毒的症状。由于未能参加晚宴，所以你自己在家做了饭，但由于一些偶然的因素，你自己做的饭让自己也食物中毒了。自己在家做饭是因为出租车开得太慢，而食物中毒是由于自己做饭导致的。

再举一个更加实际的案例：在一场凶杀案中，受害者受了重伤，但他在接受治疗时没能得到医务人员足够的重视，因此而死掉了。虽然凶杀案是导致他需要医治的原因，但在一些极端的情况下（比如医生们的行为严重违背了常规处理流程以及医疗护理“明显有误”），有人认为治疗过程才是导致死亡的原因。

1956 年，英国出现了一起非同寻常的案件。在一起谋杀案中，一名受害者被刺伤了，但受害者的死亡并不是受伤导致的，而是医疗过程导致的。[28]最终被告的谋杀罪名以及死刑判决被成功推翻。在这个案件中，受害者在被刺伤后，他的处境已经得到了改善，病情也稳定了下来。为了防止感染，医生让他服用了抗生素。之后，他出现了过敏反应，于是医生就让他停止服用抗生素。但是，另一名医生却无视他之前的过敏反应，让他重新开始服用抗生素。在对受害者进行尸检后发现，受害者是因为服用了

令他过敏的药物，并且静脉液体过多引起了肺部液体过多才致死的。因此，人们认为对受害者的治疗过程打破了受害者从受伤到死亡之间的因果关系链。[29]

另一方面，近因并不一定非要在邻近结果之前才出现，只要它能与结果的发生清晰地连在一起就行。罗纳德·里根总统的新闻秘书叫 James Brady，人们在他死后对他进行了尸检，结果发现他的死亡实际上由一起凶杀案导致的，因为他在 30 多年前的一起凶杀案中中过枪。这就是一起被滞后的凶杀案。在这种凶杀案中，受害者会在很长一段滞后期后才由于受伤而死亡。[30]在这个案例中，30 多年的时间间隔让近因暂时变成了遥远的原因，但由于有证据显示枪伤才是真正致死，所以验尸员将他的死亡判定为凶杀致死。

8.5.3 陪审团

在日常生活中，当我们想要解释一些事件的时候，可以寻找新的信息来支持或者否定我们的假设。比如说，你可以去咨询尽可能多的专家，问问他们你隔壁房子过于花哨的装饰是否会降低你房子的市值。你可以审查每一个专家的资格信息、阅读关于房价的研究报告并且进行一些实验，等等。而陪审团成员面对的则是一组他们无法控制来源的事实。在某些案件中，陪审员也许能够向证人提一些问题，[31]但在绝大部分情况下，他们只能评估并整合证据，而不能直接获得证据。

所有这些复杂的证据信息可能都不是按照时间顺序提供给陪审员的。面对这样的信息，陪审员如何才能将它们结合在一起来搞清楚究竟发生了什么事情呢？

陪审员在听审的过程中不会把每一条新的证据都放进互不相干的证据库中，以便最后一次性对所有证据做一个评估；也不会在每一个时间点对已有的证据做一个总结，然后记一份被告有罪或无罪的流水账。[32]大部

分人都认为陪审员在庭审过程中会将获取的信息组织成一个故事。这种故事模型论最早由 Nancy Pennigton 和 Reid Hastie 提出，讲的是陪审员们会将庭审提供给他们的证据（以及他们对证据的评估）和他们的先验知识以及经验结合在一起，组织出一个故事来再现案情。陪审员们会对同一个案件得出不同的结论这可能是由于他们构建了不同的案情故事。这一点正如 Reid 和 Hastie 在一次实验中发现的情况一样。[33]

对一名陪审员来说，什么样的故事才是可信的？这个问题一部分取决于陪审员的经验，另一部分取决于这个故事对证据的解释力（这个故事到底能够解释多少证据）。一名陪审员对其构建的故事的信心取决于三个关键因素：故事的覆盖面、连贯性和独特性。如果某个人有一份确凿的不在场证明，那么那些认为这个人在案件中起着必不可少的作用的故事就不可信了。因为这些故事无法解释这份表明这个人无罪的证据。这就是一个故事的覆盖面问题。同理，一个故事必须以一种连贯的方式组成一个整体。如果一名陪审员发现一名案件侦查员不可能会篡改证据，或者案件侦查员篡改证据的这个假设与故事的其余部分相矛盾（在故事的其余部分，案件侦查员没有任何篡改证据的动机），那么那些具有这种特征的故事就不是连贯的故事。在某些情况下，可能会出现多个与证据相一致的可能发生的故事。如果很多故事都是连贯的，那么陪审员就无法确定哪一种解释最有可能发生。相反，如果有一个独一无二的、连贯的并且覆盖面很广的故事，那么他们很有可能会用这个故事来解释整个案情。

然而，这并不意味着所有陪审员构建的故事都是一样的，也不意味着他们会接受同一个故事。一个陪审员相信的事，也许另一个陪审员并不相信。如果我曾有过直接经验，发现有学生在很不重要的家庭作业中作弊，却又声称自己是无辜的，那我可能就会构建出这样一个和这个学生自己的证词相矛盾的故事：该生在作弊的问题上撒谎了。相反，那些没有这种经验的人可能会发现这个故事很不可信：怎么会有学生在做对成绩影响极小

的家庭作业时作弊呢？而且他们在构建故事的过程中可能还会给该生的证词赋予更大的权重。[34]

庭审最具挑战性的特征之一是：证据是随着时间而逐条提供给陪审员的，却不一定是按照时间顺序提供给陪审员的。[35]因此，一名陪审员一开始构建的故事可能是这个学生没有作弊，他的家庭作业是在他不知情的情况下被别人抄袭了。后来有几个新的证人说看到他参与了作弊。但是，必须要将这个新的信息加入到一开始构建的故事中去。而且由于很多证据都不是相互独立的，问题就变得更加复杂了。如果我们相信证人的话而不相信该生说自己没有参与作弊的证词，那么该生其他证词的可信度可能也会降低。[36]

很多关于陪审团如何思考问题的实验性证据来自于人们针对模拟陪审团的大量研究。[37]但是，这些模拟活动可能没有抓住真实审判的一些重要特征。在真实审判中，陪审团可能会被时间跨度很长的信息搞得焦头烂额，也可能会因为案情重大（比如陪审团的任务是判定一个真实的人是否该判死刑，而模拟法庭是做一些没有任何实际影响的决定）而有不同的行为表现。除此之外，陪审团成员的选择流程本身可能也会导致真实案件中的陪审团成员与模拟法庭中陪审团成员来自于完全不同的人群。

然而，真实陪审团的审议过程一般是不公开的。[38]但亚利桑那州录影项目是个例外，这个项目记录了多个完整的审判过程，并且对这些庭审过程进行了分析，其中就包括陪审团的审议过程。[39]研究人员发现，在他们研究的 50 起民事案件中，陪审员确实针对这些证据构建了不同的故事。他们有时通过讨论共同构建一个故事，有时则是在对证据进行评估的时候，对彼此的故事提出质疑。[40]下面是陪审员们在一次庭审过程中的某一次讨论的部分内容，此时案件的证据尚未完全提交给陪审团。[41]

陪审员 1 号：他（原告）说他是在看到黄灯时提速的，提速之后灯才变红的。这一点我没弄明白——（原告）看到（被告）闯的是黄灯还是红灯呢？

陪审员 7 号：闯的是红灯，但他不得不往前开，因为他被困在路中间了。

陪审员 1 号：但是还有一次，他（原告）说他知道另外一个人看到交通灯的颜色在变，所以他（被告）提速了，也有可能那是（另一个证人）告诉他的话。那里并没有左转箭头。

陪审员 7 号：如果你看到有人加速，你会怎么办？我就坐在那里。

陪审员 1 号：对呀。

陪审员 6 号：所以我们要看看法官怎么说……这个州的法律条文是怎么规定的？

陪审员 1 号：对，车辆不应该在十字路口停留……

陪审员 6 号：但是没有打转向灯，对吗？没有箭头？那他在十字路口干什么？

陪审员 7 号：我们需要有证人来告诉我们他有没有闯红灯。

在这个讨论中，陪审员想要搞清楚一起交通事故中各个事件发生的先后顺序。他们不确定当时的交通指示灯是红色还是黄色。陪审员 7 号用一个事实（灯是红色的）和一个解释（由于被告身处十字路口，所以他不得不继续往前开）来阐明了这个问题。陪审员们对原告证词的可信度提出质疑，因为他的证词似乎发生了改变；陪审员们质疑他证词的内容是他直接观察到的还是听别人说的；然后陪审员们又将这些故事和他们自身的生活经验结合了起来。最后，他们讨论了还需要什么证据才能弄明白证人的证词。

虽然这和我们在日常生活中解释一些事件的方法是一样的，但不同的是，他们审查每一条证据的严密程度和这些证据的结合程度。不过，人们在提出各种阴谋理论时，通常会主动忽略那些相互矛盾的信息，他们一方面寻找能够证实他们理论的证据，另一方面又试图将证据纳入这些理论之中。这个审判过程为我们提供了一个解释各种事件的框架：首先要为一个原因寻找无罪和有罪的证据，然后严格审查现有的证据以确定事情的真相，最后判定到底是只有一个可信的解释还是有多个可信的解释。

注释

1. 关于这个故事的详细描述，参见 Vlahos（2012）。
2. 参见 Lange（2013）。
3. 想要了解历史上关于因果关系解释的更多内容，参见 Scriven（1966）。
4. 1938 年糕点战争是由于墨西哥一家法国糕点店遭到破坏而引起的。
5. 想要了解更多关于这方面的信息，参见 Hausman（2005）。想要了解人们关于这些困难的一些讨论，参见 Hitchcock（1995）。
6. 想要回顾一下这方面的内容，参见 Sloman 和 Lagnado（2015）。
7. Mandel（2003）。
8. 想要了解更多这方面的例子和实验，参见 Spellman 和 Kincannon（2001）。
9. Cooke（2009）；Cooke 和 Cowling（2006）。
10. Lewis（2000）。
11. 很多研究表明受训的运动员有此反应，但也有人证实那些原本不活动的参与者在参加一个锻炼项目之后，也会出现这一效应；想要了解这方面的例子，参见 Tulppo 等（2003）。
12. 想要了解相反的观点——认为主体性在这里指的是一个特征而不是一个漏洞的观点，参见 Halpern 和 Hitchcock（2010）。
13. Dalakas（1995）。
14. 想要了解医学领域关于这种不确定性的研究，参见 Hripcsak 等（2009）。
15. 想要进一步了解关于这一方法的讨论，参见 Kleinberg（2012）。
16. 这一想法被称为连接原则，是由 Sober 和 Papineau（1986）提出的。
17. 关于这一方法更全面的介绍，参见 Kleinberg（2012）。
18. 关于概率轨迹的各种讨论，参见 Eells（1991）。

19. Eells（1991）还定义了另外两种关系。当概率不发生任何变化时，结果独立于原因；当概率先变大又变小时（比如第一次接触流感的案例），结果自主发生。
20. 大部分方法都将注意力放在高级算法上，而不是方法的具体运用和实施上。Dash 等（2013）是一个例外。
21. 想要了解这个经典文本，参见 Hart 和 Honoré（1985）。
22. 桑德斯系统伯明翰有限公司诉亚当斯（1928）。
23. 想要深入了解人们关于这个案例和相关法律的讨论，参见 Wright（2007）。
24. Fischer（2006）。想要了解更多关于直觉与法律判断的内容，也参见 Fischer（1992）。
25. 想要深入了解人们关于这种案例的讨论，参见 Spellman 和 Kincannon（2001）。该文还提供了不同陪审团需要遵守的规则实例。
26. 想要了解 NESS 方法的一些问题，参见 Fumerton 和 Kress（2001）。
27. 想要了解更多信息，参见 Carpenter（1932）；Wright（1987）。
28. Rv. Jordan（1956）。
29. 注意，关于这个案例还有些争议，而且还有观点认为这个案子判得不公正。参见 White（2013）。
30. Lin 和 Gill（2009）。
31. 想要回顾这一做法，参见 Mott（2003）。
32. Lopes（1993）。
33. Pennington 和 Hastie（1992）。
34. 想要了解这一情况在 O. J. Simpson 案中是如何起作用的，参见 Hastie 和 Pennington（1996）。
35. 想要了解更多关于证据呈现顺序的影响，参见 Pennington 和 Hastie（1988）。
36. 模拟陪审团实验显示，陪审团会根据相互联系的证据而拒绝采信一些证据（Lagnado 和 Harvey，2008）。
37. Devine 等（2001）。
38. 想要回顾关于真实陪审团的一些研究，参见 Diamond 和 Rose（2005）。
39. Diamond 等（2003）。
40. 想要了解更多陪审团得出的案情真相，参见 Conley 和 Conley（2009）。
41. Diamond 等（2003），38。

第 9 章　行动

如何根据原因进行决策？

2008 年，纽约市通过了一项法案，要求拥有 15 家以上分店的连锁餐厅必须在菜单上显著标出每种食物的热量值。这项法案背后的依据是，食用高热量食物会导致肥胖症和其他健康问题，而餐厅不同于包装食品的生产商，他们往往不会在菜单上标明所售食物的营养成分信息。如果人们知道他们所食用的食物包含多少热量的话，就会改变自己的行为。然而，类似的政策在全国推广了以后，人们在纽约和其他城市展开了研究，但几乎没有发现能够表明这些法案有这种效果的证据。[1]

为什么会这样呢？在菜单上标出热量值的政策假设人们会注意到热量信息，假设在没有热量信息的情况下人们会低估食物的热量值，假设人们知道如何解读和使用热量信息，并且假设这个政策在各种类型的连锁餐厅都会产生同样的效果。这项政策不仅没有大大降低人们购买的食物的热量值，而且在某些情况下，人们购买的食物的平均热量值甚至比以前更高。[2] 比如，人们在节食或者评估不健康食品时往往会高估某些食物的热量值，[3] 而在食物所含真实热量信息公布之后，人们会觉得很惊喜，从而去点一些热量更高的食物。

如果人们不知道该如何使用这些热量数据，那他们消耗的食物的热量值可能就会增加，或者说至少不会下降。要想让热量值信息改变人们的

行为，我们必须假设消费者能够将这个信息融入日常的饮食之中，而且能够理解每一个数据的含义。如果他们不知道每一顿饭所需的热量值大约是多少，那么菜单上提供的热量信息就没有任何意义了。在研究中，人们不仅提供了食物的热量信息，还为顾客准备了一些传单，向顾客介绍每日最佳热量摄入值的区间。但这种做法也没有对人们点的食物的热量值产生具有统计学意义的显著影响。[4] 不过，这可能是因为人们在来饭店之前已经想好要点什么了，所以在销售食物的时候为人们提供这些信息就有点为时已晚了。这种信息可能也会影响人们的行为，比如他们以后会选择去其他饭店吃饭。相反，关于停车灯体系（健康的食物用绿色图标，而最不健康的食物用红色字体标出）的研究发现，有很多证据能够表明人们的行为会因为使用了停车灯体系而发生改变。[5]

只有极少数的几项研究发现，菜单上标出食物热量信息的做法是有效果的。其中有一项研究发现，在星巴克，几乎完全是由于食物购买上的变化而导致人们点的食物的热量有些许下降（6%）。[6]这 6% 的下降幅度（平均每单所点食物的热量从 247 大卡下降到 232 大卡）绝大部分是因为购买的食物数量减少了，而不是因为购买了热量比较低的食品。然而，对于一家咖啡连锁店的顾客来说，食物可能不过是顺带购买的东西。至于这 6% 的下降幅度到底有没有意义，还要看消费者在其他饭桌上有没有将减少的这 6% 的热量补充回来才能确定。

由于不同类型的餐厅提供的食物种类和面对的顾客群体不同，而不同的顾客对餐厅又有着不同的期待，所以研究中的任何效果都不可能适用于所有类型的餐厅。即便我们发现人们的购买行为发生了变化，我们也不能立即将这种变化归功于菜单上提供的食物热量信息。相反，这可能是由于餐厅改变了他们的菜单，减少了有些食物中的热量，或者在不得不公布热量信息之前将一些食物从菜单上删掉了。[7] 虽然这可能在某种意义上意味着我们的法案通过让餐厅提供更加健康的食物选择而取得了成效，但这

也意味着我们可能高估了公布食物热量信息对消费者行为的影响。

如何才能根据原因进行决策呢？仅知道跑步可以改善心血管健康状况，并不一定意味着我们已经获得了足够的信息来决定是否要开始跑步；仅知道钠元素在有些人身上可能会引发高血压，并不足以让我们决定是否应该在整个人群中实施一项限制食物中钠含量的政策。在理想情况下，我们会进行明确并严谨的实验，并在此基础上决定采取什么样的行为；但在实际生活中，我们需要在信息不完整也不完美的情况下采取行动。在某些情况下，我们根本无法进行实验，在另一些情况下，我们可能没有时间或资源等到有了明确的结论再采取行动。

但是，并不是所有信息都是同等重要的。在这一章，我们会将之前讨论过的理论综合在一起，形成一组需要注意的事项，以此来评估各种因果假设。我们将会讨论我们到底需要什么样的信息来支持某个因果假设，以及有什么好的证据能够证明某个因果关系中含有这些特征。虽然因果关系的一个基本特征就是原因能够提高某个结果发生的概率，但是呈现这一特征的方式也各不相同，这就有可能让人们得出完全不同的结论。找到原因只是第一步，要想针对整个人群以及每个个体成功地制定一些政策，我们还需要更多的信息。当我们决定采取行动时，无论是通过改变饭店的标志来改善顾客的健康状况，还是选择一种药物来缓解头疼症状，我们所做的都不仅是在确定是否要去做某个具体的事情，而是在能够导致同一结果的很多方法中做出选择。一个原因可能在一个地方有效，但在另一个地方没有任何效果，或者可能会导致副作用（既有积极的副作用也有消极的副作用）。所以，我们将讨论如何预测干预措施的效果，以便做出更好的选择。此外，并不是所有的原因都能经得起干预措施的考验，而且干预措施让一个原因出现的同时可能还会改变其他一些事物。所以，我们将考察为何

需要考虑要使用哪一个原因来引发某种结果（比如公布食物的热量值或者强制要求公布食物的热量值），还会考察如何让某种结果出现（比如对公布食物热量值的饭店给予奖励，或者对不公布食物热量值的饭店予以处罚），以及当结果出现时，还有其他什么可能会因此而发生改变的事情（比如餐厅更改了菜单，低热量值的甜味剂的消耗量增加了）。

9.1 对因果假设的评估

没有任何一个测试因果关系的方法能够适用于所有的情况，但在面对实际问题时，我们仍需要做出因果假设并对其进行评估。电视剧《十六岁的怀孕女孩》真的像该剧宣传的那样能够降低播放该剧的地区青少年怀孕的比例吗？[8] 我们没有随机挑选任何人来观看这部电视剧，而且在大部分情况下，我们甚至都不知道某些人有没有观看这部电视剧。从理论上来讲，我们可以随机分派一些青少年去观看不同的电视节目，但由于现实中青少年怀孕的情况十分少见，所以我们根本找不到足够大的样本库来观察观看这部电视剧的效果。

我们已经讨论过如何成功使用随机试验去寻找事件发生的原因，但在很多情况下，我们都做不了这样的试验。这时，我们需要对其他证据进行评估，以此来确定某种关系是因果关系的可能性。不仅如此，我们从理想并完美的随机试验中了解到的信息与我们从任何给定的真实试验中了解到的信息也是不同的。真实的试验可能并非盲法试验，试验的样本可能会很小，而且在试验的过程中，很多参与者可能在试验还未结束时就已经退出了。

RCT 在任何情况下都比观察性研究要好，这种说法是不准确的。[9] 对于一个要在不同治疗方案中做出选择的病人来说，如果一边是一项大规模、长期且针对与她的症状完全一样的病人群体的观察性研究，而另一边是一

个规模很小的并且针对与她的病症不同的男性病人（而且她已经试过其他几种治疗方案，都没有见到任何效果，而这些男性病人还没有试过其他几种治疗方案）的随机试验，那么在这种情况下，前者可能会为她的决策提供更好的证据。这正是我们在第 7 章讨论过的外部有效性问题。如果 RCT 并不适用于我们将要干预的情况，那么这个试验就不是针对这种情况来采取干预措施的最佳证据。即使事件发生的背景是一样的，观察性研究（这种研究可能会重新使用现有数据，比如电子病历数据）能够做的事情与 RCT 能够做的事情可能也有所不同。如果我们想知道坚持不懈地锻炼几十年会对人们的衰老过程产生怎样的影响，以此来指导我们制定当下的公共政策，那么一边是过去 50 年来对数万人的观察性研究，另一边是针对 100 名参与者进行的为期两年的 RCT，两者相比，前者的指导效果可能要更好。尽管 RCT 常被当作衡量因果假设证据的黄金标准，但即使没有实验研究，我们依然可以掌握事件发生的原因。因此，我们有必要知道如何去评估非实验性证据。[10]

---

20 世纪 60 年代，Bradford Hill 提出了一组在评估因果假设时需要考虑的因素。[11] 这些因素有时会被误认为是验证因果关系的一组标准或者一个清单。虽然这些因素中的任何一个因素都不是必要条件（即便不是所有因素都出现了，事物之间也仍有可能会存在因果关系），而且整个因素组合也不是充分条件（即使整个因素组合中的所有因素都出现了，这个案例中的因果关系可能也是虚假的），但在我们无法进行实验研究的时候，这个因素组合仍然可以为我们提供一些需要考虑的因素，并且能够将我们前面讨论过的很多理论结合在一起。[12]

组合中的因素大致可以分为两种类型：第一种类型的因素为我们指明某个原因对结果产生了影响（强度、一致性以及生物梯度），第二种类

型的因素则为我们提供了证据，这些证据表明存在着某种可以让某个原因对结果产生上述影响（特异性、时间性、可信度、连贯性、实验以及类比性）的机制。尽管这个影响因素的清单与 Hill 提出的需要考虑的因素在顺序上并不完全一致，但下面的内容里还是保留了这个顺序，主要是为了方便我们交叉参考关于这些因素的相关论文。[13] 在对这些因素进行评估时，我们会考察其中的每一个因素，同时也会提出一些需要进一步思考的问题。

9.1.1　强度

如果在菜单上标出食物的热量值可以降低人们在点餐时选的食物的热量值，那么人们在标出食物热量值的餐厅里所点的食物的热量值应该明显低于他们在那些没有标出食物热量值的餐厅里所点的食物的热量值。同理，有些地方播放了有关青少年怀孕问题的电视节目，还有一些地方没有播放这样的电视节目，如果前面那些地区的青少年怀孕率只比后面那些地区略微低了一点点，那么用这种数据来证明那些电视节目可以改变青少年的怀孕率就没什么说服力了。相反，如果在上述两个案例中，无论是人们所食用食物的热量值还是青少年的怀孕率都显著下降了，那么这样的数据就能够更加有力地证明事物之间的因果联系。这与因果概率法（详见第 5 章）密切相关，因果概率法研究的就是在某个原因出现后，某种结果出现的概率提高的幅度。这种方法还和第 6 章讨论的测量因果关系强度的方法有着十分紧密的联系。强度可以指让一个事件发生的可能性更大（公布食物热量信息极大提高了人们购买低热量食物的概率），也可以指让某种影响的力度更大（公布食物热量信息导致人们购买的食物热量降低了一半）。

然而，事物之间的联系不强并不意味着它们之间就没有因果关系。因为有些原因可能会比较弱，比如吸二手烟导致肺癌的比例要比吸烟导致肺癌的比例小得多。还有一种原因很弱，但是仍然会对结果产生决定性的影响：所有遵循某个节食计划的人，体重都有所下降，但他们减掉的重量

只占其原体重的很小一部分。还有可能会存在一些我们尚未发现的更小的群体，比如公布食物热量信息只对那些已经在计算食物热量的人群有效。在这种情况下，如果我们把所有人的数据结合在一起来分析，那么事物之间的联系可能就显得微不足道了。

我们还讨论过很多这样的情况：事物之间可能会出现很强的相关性，却不存在相应的因果关系。唐氏综合征和出生顺序就是这样一个例子。出生顺序可以向我们透露母亲生育孩子时的年龄（生第四个孩子的母亲平均要比生第一个孩子的母亲的年龄更大），因此出生顺序和唐氏综合征之间有着很强的相关性，但它却不是导致唐氏综合征的原因。[14]相关性的强度是否能够有力证明事物之间存在因果关系，这要看我们是否解释了这些可能的共同原因，以及这些共同原因是否能够解释各种结果之间的联系。[15]

当我们看到事物之间存在很强的相关性时，我们需要考虑的问题有：这种关系是不对称的吗（为什么我们会认为其中一个事物是原因而另一个事物是结果呢）？这种相关性是否是这两个事物之间的一个共同原因导致的？这种相关性是否是方法问题（范围限制、选择性偏差和失误）导致的？我们是否忽略了其他与结果密切相关的因素？对于那些时间序列数据来说，这种相关性是否是两个事物都是非稳定变量导致的（也许这两个事物都随着时间的变化而呈现出一种相似的上升趋势）？

9.1.2 一致性（可重复性）

如果公布食物热量信息确实可以降低人们摄入的热量值，那么不同的研究人员通过不同的方法应该可以重复获得这一发现，而且这一发现应该在多家餐厅都适用。虽然这与 Hume 和 Mackie 理论中的规律性并不是同一个概念，但两者的思路是一样的——真正的因果关系不应该只能在一个试验中观察到，而应该在很多试验中都能观察到。我们在第 7 章讨论过，出于很多原因，有些发现可能是无法复制的。但是，针对很多城市的不同

人群，不同的研究人员使用不同的方法对公布食物热量信息的影响进行了研究，然后发现公布食物热量信息并未降低人们点的食物的热量值。这么多的研究让这一发现成为偶然性事件的概率大大降低。在重复试验的过程中，我们所引进的变量会很自然地导致我们对事物之间关系的强度得出更加肯定的结论。然而，我们在某一个城市发现公布食物的热量信息导致人们在几家咖啡店点的食物的热量值有所下降。

我们也可以用结论的不一致性来排除一些表面上很密切的因果关系。有很多论文研究了哪些食物似乎可以增加人们患癌症的风险，以及哪些食物似乎可以降低这些风险。通过分析这些论文，我们发现几乎每一种食物都有能够增加或降低人们患癌症的风险的证据。[16] 人们可能会从他们所引用的研究中挑出那些支持自己想法的有力证据，但在考察这些研究的所有证据后，我们发现这些证据并不是那么确凿。同理，由于一次性检测很多假设（所以某个假设可能由于巧合而具有了显著性）所导致的假阳性结论也是不可重复的。

当我们的发现不具有一致性时，又能得出什么样的结论呢？有可能让某个原因起作用的关键因素在一个试验中出现了，但在另一个试验中却没有出现。比如说，被很多蚊子叮咬并不一定会得疟疾，只有感染疟疾的蚊子才会传播疟疾。如果我们不知道起作用的关键因素是什么，那么结果似乎就变得无法预测了。值得注意的是，研究结论不一致并不等于原因本身不一致。正如疟疾的案例一样，可能是由于我们研究的群体在一些关键问题上存在差异，所以导致了研究结论的不一致。

在所有研究中都一致的发现也有可能是一个共同的缺陷或疏忽导致的。如果每一项研究都只记录了出生顺序而没有记下产妇的年龄，而产妇的年龄实际上能够准确地反映出生顺序，那么虽然出生顺序和唐氏综合征之间没有因果关系，但两者之间的联系仍然会呈现出一致性特征。同理，有可能所有的研究都犯了相同的数学错误或者都使用了同样被污染的样本。

在评估某个关系的一致性时，我们需要考虑的问题包括[17]：我们是否准确复制了那些研究方法？研究的目的是再现主要结果吗？如果我们未能成功复制一项研究，这是否可能是研究群体或研究方法的显著变化造成的？在不同的研究中，结果的大小是一致的吗？这些研究都有足够的动力可以让我们发现某个原因导致的结果吗？这些研究是彼此独立的吗（或者资金来源是否相同，比如一家制药公司同时资助两项研究）？

9.1.3 特异性

如果有人说单独服用某一种药品能够治好癌症、普通感冒和疟疾，我们肯定会认为这种说法十分不可信。但是，我们却知道吸烟会在不同程度上导致很多健康问题。

特异性指的不仅仅是一个原因导致的各种结果之间的差异，还包括这个原因对每一个结果的影响程度。这并不意味着一个原因只能导致一种结果（这也不大可能），而是意味着与一个似乎对每种结果都会产生重要影响的原因相比，一个更加具体的关系可能会为我们提供更加强有力的证据。比如说，某种药物可能无法完全治愈很多不同的疾病，但它却可能对某一种疾病产生主要效果，而对其他疾病产生次要效果。同理，如果有人说骑行可以减少各种原因导致的死亡事件，那么这种说法似乎也是令人难以置信的。然而，如果我们说骑行对于健康的主要作用是可以减少肥胖症患者的数量以及心血管疾病导致的死亡事件，那么这种说法就比较可信了。

从某种意义上来说，特异性还意味着我们推理出的关系到底有多直接。在特异性的一端，我们可能会看到粒度非常细的关系，比如说我们发现周三早上发出的竞选募捐邮件与周六晚上发出的竞选募捐邮件相比，能从收件人那里筹集到更多、额度更大的捐款。而在特异性的另一端，我们可能只会发现筹集到更多资金与候选人发邮件有关。

特异性往往取决于我们的认知程度。如果我们对某个原因的作用机制以及它的主要影响一无所知，那么我们得到的可能只是（反映事物间关系的）非常间接的证据（比如只考察吸烟者的死亡率与肺癌发病率及死亡率）。虽然特异性并不是必要条件，但与事物之间的间接关系相比，人们可能更容易接受事物之间更加密切的直接关系。不过，人们一般认为特异性是一个相对来说不太重要的标准之一。[18]

至于有没有可能产生多重效应，这要取决于我们假设的关系的运行机制。假如我们认为因为自行车头盔能够降低骑行者头部受伤的概率，所以头盔对骑行者具有保护作用。这种情况下，如果我们说戴头盔能减少骑行者各种类型的受伤事件，或者戴头盔能减少骑行者中头部受伤的事件，而对其他类型的受伤事件影响极小，两者相比，后者能够更加有力地证明头盔对骑行者的保护作用。这是因为总的受伤事件的减少可能是因为戴头盔的骑行者骑车更为谨慎或更有经验，而这些人与不戴头盔的人相比，受伤的可能性本来就更低。[19]

因此，在考虑特异性的过程中，我们还必须考虑事物之间联系的强度以及我们的先验知识：这个原因会导致不同的结果吗？它对各种结果的影响程度是一致的吗？这个原因对结果的影响程度与我们预期的影响程度是否有差别？

9.1.4　时间性

是青少年的怀孕率下降导致观看反映青少年怀孕问题的电视节目的人数增加，还是观看这种电视节目的人数增加导致青少年的怀孕率下降？我们在第 4 章讨论过，事件发生的顺序也是寻找因果关系的一个重要线索。但有时我们并不知道哪个在前、哪个在后：是一通电话改变了选民们的投票偏向，还是因为针对选民的人口统计学数据的分析预测到了这些选民的偏向，所以他们的名字才会出现在需要打电话游说的选民名单中？理

清事件发生的顺序是弄清因果关系真实方向的关键。

比如说，某种疾病的早期症状可能会出现在这个疾病被确诊之前，但实际上是这种疾病引起了这些症状。在随机试验中，干预措施和干预结果的顺序是清晰的，我们可以从观察性时间序列数据中发现这种顺序（假设测量的频率足够高，这样如果 A 出现在 B 之前的话，那么这两件事就一定会先后按顺序被观察到）。但有些研究使用的是一次性案例，这些研究在面对这个问题时可能会遇到一些麻烦。这些横断面研究就像是给研究群体拍了个快照，比如调查人们的居住地址以及有什么过敏问题等。但是，这样的研究只能告诉我们某一次出现了什么情况，我们无法知道这些人在搬到某个特定的地方之前有没有过敏史，也不知道他们的过敏问题是否是新的环境导致的。

尽管时间上的优先性意味着原因会在结果之前出现，但我们也必须考虑原因和结果之间隔了多长时间。我们是否会相信原因和结果之间会出现一个很长的时间间隔，这取决于我们已经掌握的信息。如果你看到有人进入一个很陡峭的封闭式滑道，你一定会认为他出现在滑道底部的速度比在平缓一些的滑道中要快得多。所以，在第一种情况下，耽搁很长时间是不太可能的；而在第二种情况下，耽搁的时间很短也是不太可能的。我们在第 4 章心理学研究中已经见识过这一点了。在第 4 章的一个实验中，当原因和结果之间的时间间隔很短时，参与者们认为存在某种因果关系的可能性更大。只有当他们知道其中的作用机制需要更长的运行时间时，他们才会在原因和结果之间的时间间隔较长时也认为存在某种因果关系。人们很难相信在接触石棉和患上癌症之间只有几分钟的时间间隔，但人们却很有可能会在看到食物热量信息后的几分钟内就改变他们所点的食物。

即使原因确实发生在结果之前，它也不一定是当时唯一发生的事。如果公布食物热量信息和餐厅对菜单进行巨大改动这两件事同时发生，那我们就无法确定让顾客改变行为的是哪一件事。比如说，有些研究曾经认为

一个人的小学老师会对这个人几十年后的工资产生影响。[20]为了证明这是可信的，我们必须找到一个原因来解释在这几十年中存在一个从童年就一直延续下来的影响（最终导致了一些与工资有关的其他事件链），而且这个影响没有被某个共同的原因混淆，也不是一些其他的中间原因导致的。

无论我们是否看到原因出现在结果之前，都必须考虑一些问题：这些事件之间的表面顺序是正确的吗？这是否是一个由数据收集方式或失误导致的人为结果？考虑到原因的运行机制，这种时间间隔合理吗？在假设的原因出现之后存在一个很长的时间间隔，那么这个结果有没有可能是其他因素的干预导致的？反之，在有可能导致结果的原因出现的时候，还有没有其他几乎在同一时间发生的事件呢？

9.1.5 生物梯度

是不是越多的原因就会导致越多的结果呢？这正是 Mill 的共变法研究的问题。随着原因的剂量增加，它引起的反应也应该增加。[21]随着工人们在被石棉污染的环境中待的时间越长、与石棉的接触越多，他们患上疾病的风险也应该越大。相反，人的身体对葡萄酒的反应就不会那么敏感，稍微多喝一点或者少喝一点不会有太大的差别，所以，“每天喝正好一杯葡萄酒才是唯一对身体有益的饮用量”似乎不太可信。“剂量”也有可能是建立在距离基础之上的，比如在 Snow 发现霍乱原因的案例中，伦敦居民居住的地方距离被污染水泵的远近。[22]如果说在一个巨大的半径范围内，所有人得霍乱的风险都是完全一样的，或者随着人们距被污染水泵越远，患上霍乱的风险就越小，二者相比，第一种说法作为证据的说服力显然要小得多。

此外，如果一个人接触原因的情况发生了变化（比如停止服用某种药物、戒烟或者减少钠的摄入量），那么那些由于接触导致的副作用、患癌症的风险以及高血压也应该会发生变化——假设这种影响不是永久不变的。

不过，Mill 的方法需要注意的问题在这儿也同样需要注意。比如那个酒精和心脏疾病的案例，当酒精摄入量很高或很低时，患心脏病的风险都比较高；而摄入量不高不低时，患心脏病的风险会降低。很多生物性关系都会呈现这种 J 形曲线（如图 5-1 所示）。在这种曲线中，剂量低的一端风险更高，剂量中等时风险下降，然后在剂量高的时候风险又迅速回升。

我们需要主要考虑的一些问题包括：针对不同的原因值，结果的量（或出现的可能性）会发生怎样的改变？如果我们能够控制一个人与原因的接触，这是否能够改变那个人所面临的风险程度？或者是否能够改变原因所导致的各种结果？我们对剂量的测量到底有多精确？

9.1.6 可信度与连贯性

根据我们当下掌握的知识，是否可能存在一种能将原因和结果连接在一起的机制？[23] 如果我们提出咖啡饮用过量会导致人们英年早逝，在这种情况下，如果我们知道这种结果是如何出现的，而且我们的解释与当下人们对生物学的理解是一致的，那么这种说法将会更为可信。如果太多的咖啡因会让人们紧张不安并且降低他们对正在执行的任务的意识，那么他们就很可能会陷入更多的事故之中。相反，如果我们提出总统穿暖色衣服时股市就会上涨，穿冷色衣服时股票价格就会急剧下跌，那这就需要我们从了解的股票知识跨越一个巨大的鸿沟到新的结论。

由于我们的认知可能是错误的，而且可能并不知道一个新原因的作用原理是什么，所以 Hill 认为可信度并不是必不可少的东西。然而，我们要有一个通过原因产生结果的假设机制，而且其他研究人员已经强调过这种假设机制的重要性了。[24] 我们可能最终并不需要这种证据，但它却能让我们对自己的发现更加自信。关系越古怪，我们就越需要这种信息作为支撑。

根据我们当下掌握的知识，这种可能存在的关系具有连贯性吗？这个关系和我们通常认可的事实是否矛盾？它和我们的认知一致吗？由于

我们的知识也可能是错误的，所以这并不是一个无法跨越的障碍。然而，如果一个可能存在的因果关系和我们掌握的物理学知识矛盾（包括万有引力），那么我们就需要三思而后行了。[25]

要注意连贯性和可信度之间的差别。可信度指的是根据我们掌握的知识，我们能够想到某种方式来让我们研究的因果关系得以出现。但对于连贯性来说，我们可能对原因导致结果的方式一无所知，但当原因和结果联系在一起时，却与我们的认知并不矛盾。当 Snow 第一次发现被污染水泵和霍乱之间存在联系时，人们根本不会想到导致霍乱暴发的竟然是被污染水体中那些微小的细菌。当时人们都认为霍乱是被污染的空气导致的，Snow 的发现与人们的认识格格不入。随着时间的变化，我们掌握的知识也会发生变化，因此我们对于“什么是连贯的”“什么是可能的”的看法也会发生变化。

所以，当我们评估某种关系是否可信或者是否连贯时，也必须评估自己已有的认知。如果这种新的关系与我们的认知矛盾，我们又有多大把握保证我们的认知是正确的?

9.1.7　实验

如果我们通过干预措施来引入导致结果出现的原因或者提高原因出现的概率，那结果会出现吗?这个因素和其他因素之间最大的差别在于，它要求我们积极地操控某个事物，而其他因素则完全可以通过观察得到。然而，实验也不一定非要是在人身上进行的随机对照实验。在有些情况下，这样的随机对照实验也许是不可能的、不可行的或者需要的时间太长以致我们无法得出结论，所以实验结论也可能来自于体外研究或者在动物身上进行的实验。虽然我们从未做过强迫人们去吸烟的实验，但是有实验表明，将焦油涂在动物耳朵上会导致那个部位发生癌变，这就为我们提供了辅助性证据，证明烟草中的某种成分有可能是致癌物质。实验研究让我们能够

切断引起我们采取干预措施的事物和其带来的结果之间的联系。因此，如果在一个虚假的原因和结果之间存在一个共同的原因，那么对虚假原因的操纵将不会对结果产生任何影响。

我们在第 7 章已经讨论过很多原因，它们既有可能让我们在实验中无法找到真正的因果关系（比如在样本规模太小的情况下），也有可能让我们找到的原因是一个虚假的原因（比如在非盲随机试验中）。在以动物为实验对象的研究中，即使研究结论是阳性的，我们也必须仔细考量手中的证据，以便保证实验研究的原因在我们研究的系统中与在人群中的表现是一致的。比如我们以老鼠为研究对象，发现了一些治疗败血症的方法。这些治疗方法本来在人身上也应该起作用，但是我们却没有发现这样的证据。结果就有人提出质疑：在研究人类的各种炎症类疾病时，用老鼠作为实验对象是不是一个好的选择？[26]

当我们不在人身上做实验或者在活体以外的环境中做实验时，必须要确定所用的替代品是否能够反映某个原因在人体中起作用的方式。

9.1.8 类比性

最后，如果我们了解到有一个相似的因果关系，那就可以相应降低对证据的要求，因为这个相似的因果关系已经证明了某个原因是有可能导致我们想要证明的结果的。假设我们了解到，如果餐厅标出食物的脂肪含量数据，那么顾客点的食物的脂肪含量就会下降。由于我们已经知道公布食物的营养数据有可能会改变人们的行为，所以我们就更有可能会相信公布食物的热量信息有可能改变人们的行为。再打个比方，在我们知道乳头瘤病毒会导致一些子宫癌之后，我们会发现一种病毒能够导致不同癌症的说法更加可信了。类比还意味着可以利用关于动物的研究来更好地了解人类，或者可以将不同规模的各种系统连接在一起。

我们必须要评估一个实验的装置与我们想要研究的系统之间到底有多接近，也必须要审查我们手头到底有多少证据能够证明我们从一个环境中了解到的信息能够应用于另一个环境之中。

记住，没有任何一个清单可以确定事物之间的因果关系，也没有任何一个必须满足的或者始终能够满足的因果关系标准。上面分析的各种因素只是将概率法、机械法、干预法和实验法等方法结合在了一起，形成了一组需要考虑的因素。在每个案例中，我们都必须考虑信息本身的质量。随机实验的信息质量可能很糟糕，事物之间的相关性可能是选择性偏差导致的结果，而用动物做实验对象的研究结果可能并不适用于某种特定的疾病。同理，证据的标准也取决于这个证据要支持的观点到底是什么，以及由此导致的行为的潜在风险和成本。哲学家们曾经提出过一些关于证据的理论，试图描述某个事物作为一个科学假设的证据到底意味着什么。但一般来说，这些理论和科学家们实际使用和看待证据的方式大不相同，而且这些理论往往忽略了使用证据时的背景所起的作用。[27]

比如说，谋杀案的证据标准就比究竟是哪个小孩打碎了花瓶的证据标准要高得多。因为在第一个案例中，发生冤假错案的后果要比第二个案例严重得多。有一个很弱的证据表明，每天吃一块巧克力可以改善人们的心情。这样的证据对于一个人决定坚持吃巧克力可能是足够的，但是却不足以让我们制定一个建议人们每天都要吃巧克力的国家营养标准。

9.2　根据原因制定政策

将苏打水的瓶子变小、在连锁餐厅的菜单上公布食物的热量值、禁用反式脂肪以及降低餐厅食物的含钠量等，这些只不过是纽约市为了改善纽约人口的健康状况曾考虑或实施过的部分政策。

如果我们知道糖、高热量食物、反式脂肪和钠与我们想要改善的各

种健康问题之间存在因果关系，那么我们能预先知道纽约市的上述行为是否会取得成功吗？要想理解这个问题，我们需要知道一项干预措施的影响是什么，以及如何在各种可能的干预措施中做出选择。然而，一种行为导致的影响不一定仅仅是我们能够想到的结果。一个原因可能会导致多种结果，更令人头疼的是，干预行为本身可能也会导致事物之间的因果关系发生改变。某种降低胆固醇的药物可能在一个人身上的效果非常好，但在另一个人身上却完全无效。这是因为另一个人认为这种药物无论怎样都能帮他控制胆固醇，所以他就选择了对身体更加有害的食谱。在另一个案例中，如果标准化测试成绩一开始和教学质量紧密相连，但人们用考试成绩来评价老师，那么标准化测试成绩和教学质量之间的联系就可能会变弱。因为在这种情况下，老师会把他们的教学中心完全放在为学生备考上。[28]

尽管如此，我们仍然想把决策建立在证据之上，而不是建立在传闻轶事之上。而且，证据也应该建立在因果关系之上，而不是建立在相关性之上。现在出现了以证据为依据的医学、设计、教育以及很多其他运动来推进以证据为依据的行为。这并不是说这些领域以前不以证据为依据，而是说那些主张采用以证据为依据的研究方法的人在试图确定什么是好的证据。他们不再简单地判断某个给定的假设是否有证据作为支撑，而是试图区分有力的和无力的证据，并且推动人们去使用更好的证据。这一切的结果通常会呈现出一个等级不同的证据金字塔体系，[29]而 RCT（或者更具体地说，是对多个 RCT 的系统性综述）无一例外地会出现在这个金字塔的顶部。然而，这些等级体系并不一定会告诉我们什么样的信息是必要的，以及如何去使用这样的信息。从理论上来说，一个完美的 RCT 可能是最好的证据，但在现实生活中，我们对比的并不是一个完美的实验和一项观察性研究。相反，我们可能会面对一个规模很小且带有偏差的随机试验研究和一个规模很大又很完美的观察性研究，这两者给出的证据还是相互矛盾的，或者我们只掌握了一些非实证性证据。在实践中，我们不得不依据

这样的信息来采取行动。所以，知道如何以更好地方式来应对这种情况对我们来说至关重要。

我们将考察各种因素来决定什么时候实施某项政策以及如何得出一般性结论。这里所说的“政策”或“干预措施”可能是一项禁止在全市范围内的公共场所吸烟以改善市民健康状况的政策，也可能是美联储调整利率以刺激经济活动的政策，还可能只是一个让你不要在下午四点之后喝咖啡以便减轻失眠症状的要求。在所有这些情况中，我们都引入了一种变化以便实现某个特定的目的。在某些情况中，我们的证据可能是在一个地方实施的一项政策（比如在纽约市要求餐厅公布食物所含热量的政策），而我们想在另一个地方实施同样的政策，以便实现同样的目的。

纽约、伦敦和巴黎等城市都曾实施过公共自行车项目。用户可以在一个地方取用一辆自行车，然后在靠近目的地的地方归还这辆自行车。这个项目试图减少人们乘坐汽车出行的次数，并且通过促进人们从事更多的体力活动来改善人们的健康状况。[30]这个项目能否实现其预设目标取决于以下几个假设：(1) 骑自行车是一种有效的锻炼形式；(2) 这个项目会增加人们骑自行车的次数（而不只是让人们放弃骑自己的自行车而改骑公共自行车）。但是，我们如何才能知道这些假设是否合理？如果我们试图在另一个城市实施公共自行车项目，又会出现什么样的情况？

我们可以利用第 6 章讲过的模型来预测实施干预措施的效果。但这要假设我们所用的模型是完整且正确的，而且我们在实验或者试点研究中掌握的信息会运用到现实生活中去。在那些模型中，干预措施曾是一种非常精确的工具，它能在不改变其他变量的情况下，通过某种方式来确定一个变量为真或为假。模型通常只能在我们一次只操纵一个变量的情况下告诉我们会发生什么样的情况，但在现实生活中，我们的干预措施会带来很

多变化，而且会带来这些模型预测不到的结果。

一旦决定要推广骑行运动以便改善人们的健康状况，我们就有很多方法可以实现这个目标。我们可以赠送自行车、举办骑行培训课程以及引入自行车共享计划，等等。每一种干预措施都可能会导致不同的结果。甚至一旦我们选定一个干预目标（比如自行车共享计划），还可以通过很多方式来实施它。我们还需要考虑很多因素，比如必须要确定这个项目的资金由谁提供、自行车停在哪里以及是否应该为骑行者提供头盔（或要求骑行者戴头盔），等等。因此，我们不仅是在试图确定要使用哪一个原因来导致某种结果，还是在试图明确如何才能让这个原因出现。

9.2.1 背景

我们需要了解的首要信息之一就是一项干预措施发生的背景。是否只有在有了受保护的自行车道的情况下，这种自行车共享计划才能实现？这个计划是否需要一个足够大的、已经存在的骑行群体？这个计划是否只有在人口密度比较高并且有很多自行车停放点的城市才能实施？第 5 章介绍的 Mackie 所用的方法以及用原因组成的饼形图考察了一组条件，一个原因要想引起某种结果，还需要满足这一组条件才行。

为了成功干预，我们需要知道哪些因素能让一个原因生效，以及我们要实施某项政策的地方已经具备了这些因素。我们还需要知道那些有可能让原因无效的因素都不会出现。比如，由于某种新药的价格太高，病人未能按照要求的剂量服药，那么这种药物就不会产生效果。[31] 如果一个城市没有自行车道，而骑行者又发现在机动车道骑自行车很不安全，那么自行车共享计划可能就不会被采纳。比如，在华盛顿的哥伦比亚特区，有研究发现自行车共享计划停车点的使用频率和它是否靠近自行车道存在相关性。[32]

了解背景可以帮助我们预测一项干预措施是否会成功，并且可以帮

助我们解释为什么某项干预措施可能已经失败了。这里所说的背景指的是原因饼形图中的其他原因或者其他 INUS 条件。要想让原因能够产生某个结果，这些都是必不可少的因素。如果没有这些因素，一项实验研究可能在一个地方能够证明某个干预措施是有效的，但在另外一个地方可能就无法证实了。

蚊帐是预防疟疾的重要手段，但蚊帐的使用也面临着很多障碍，其中包括蚊帐的价格。免费发放蚊帐应该可以减少疟疾的发病率，但这种结果只有在人们按照要求使用发放的蚊帐时才会出现。虽然大部分地区的人都是按要求使用蚊帐的，但还有一些地区的人却把蚊帐用作捕鱼的渔具。因为这些地区缺乏食物，所以与疟疾相比，饥饿是一个更迫在眉睫的问题。[33]所以在实施干预措施时，我们需要有证据来证明这些蚊帐会被用来解决我们想要解决的问题，或者需要有一个政策来解决这些影响蚊帐起作用的因素。[34]

有一个问题是，如果这些因素没有被测量到，那么人们对它们的存在可能还是一无所知。如果自行车道的存在确实能够导致更多的人骑自行车（而不是将自行车停车点设置在自行车道旁边），那么在一个新的场所，如果我们不知道那里有没有自行车道，或者根本就不知道自行车道的必要性，那么这个公共自行车项目在这个新的场所就有可能会失败。

9.2.2　效力和效果

一项干预措施完全失败的情况极为少见，但在现实生活中发生的情况（效果）可能也会与人们预测的结果大不相同，因为这些预测结果都是根据理想化设置推理出来的（效力）。[35]效果和效力的差别在医学上最为明显。但是无论什么时候，只要我们使用来自控制条件下的信息来指导其他背景下的干预措施，就有必要想一想效果和效力的差别是什么。

比如，由于在日常生活中，人们不太注意血样污染和洗手的问题，

所以指尖血糖仪在现实生活中测出来的结果就没有在控制条件下测出来的结果那么准确。[36]在一项研究中，因为某种药物每天都在同一时间服用，所以效力很高；但在现实生活中，因为每天服药的时间变化幅度很大，所以它的效果可能就没有那么好了。因此，如果我们只假设某种干预措施的效果会和我们在控制条件下观察到的效果或者在研究不同人群时看到的效果一致，那我们可能会高估这种干预措施的实际效果。因为病人实际上可能不会准时服药，也可能不会按剂量服药，还可能会在疗程结束之前停止服药。

效力和效果不同的可能性的大小（以及它们之间差别的大小）会直接影响我们对干预措施的选择。我们是否有理由认为在真正实施干预措施的时候还能保持同等规模的影响？在选择不同的干预措施时，我们不仅要考察哪些措施是有效的（比如什么样的干预措施让人们点的食物所含的热量值下降了），还要考察那些有效措施的作用有多大（比如人们点的食物所含的热量值下降了多少）。如果在理想的情况下（通常就是在最好的情况下），人们点的食物所含的热量值只下降了一点点，那我们就不应该认为这项干预措施在现实生活中的影响会比在理想情况下要大。同理，我们还必须考虑影响规模的分布情况。如果人们点的食物所含的平均热量值下降得很少，那我们就要弄清楚是否在所有的情况下热量值下降的数量都是相似的，还是这个平均数掩盖了一些热量值下降幅度极大和极小的情况（人们在一个地方点的食物的热量值下降了很多，而在另一个地方点的食物的热量值下降了很少）。

实施干预措施的环境可能与发现因果关系的实验环境并不一样。了解这一点可以帮助我们预测干预过程中可能出现的失败情况，并且帮助我们提出不同的干预策略来避免出现干预失败的情况。因此，在决定选择什么样的干预措施时，不仅要考虑这个干预措施的效果如何，还要考虑这个干预措施在实际出现的条件下是否能够取得成功。

9.2.3　意外的结果

一个叫作田纳西州 STAR 项目的随机试验发现，被分到小规模班级的学生在标准化考试中的成绩比被那些分到规模比较大的班级的学生要好。[37]在这个试验中，我们知道实施小规模班级干预措施的具体细节，通过随机分配各个小组，试验考评者排除了其他因素的影响，确保不可能出现某个既会导致班级规模变小，又会导致学生考试成绩变好的因素。毕竟，班级规模比较小的学校由于种种原因可能会比其他学校做得更好，而且小规模的班级可能也只是为我们提供了一个指示器，指出这所学校里有让小规模班级学校做得更好的原因。

在加州，人们一直担心班级规模太大对学生不好，随着田纳西州 STAR 项目得出的积极结论，加州实施了一个数十亿美元的项目来缩小班级规模。[38]在田纳西州的实验中，教师和学生被随机分配到规模不一的班级中；在加州，州政府为每个学生提供了 650 美元的奖励措施，用来推进各个学校缩小班级规模的措施。这个项目很快为各个学校所采纳，但是，由于班级规模变小而学生数量不变，学校就会需要更多的老师。由于师资力量跟不上不断增长的需求，在这项政策实施后，教师队伍中无经验老师的占比上升了。[39]

在那些低收入校区和少数民族校区，由于教室的数量不够，这个政策的推行时间更长了。由于教师数量不够，这个政策又未能及时实施，这些校区一度处于劣势。结果，这些校区最终招聘到的教师中有 20%以上没有各种资格证。[40]然而，田纳西州 STAR RCT 的一个主要发现恰恰是少数民族的学生从小规模班级中受益最大。加州迅速激励所有学校来实施这个干预措施，导致加州学校对教师的需求激增，而各个学校争抢师资的结果恰恰让那些本该从这个项目中受益最大的学校落在了后面。

最终，人们并不认为这个项目是成功的。任何认为这个项目有益的

论断都是证据不确凿的，或者只在小范围内存在的。而且，人们担心这个项目进一步扩大了教育上的差距。同时，即使加州的干预措施确实在一些学校产生了一点效果，也不是没有代价的。数十亿美元的项目资金意味着这些钱没能花在其他项目上，而建新教室所需的场地也是从其他项目（比如特殊教育、计算机实验室和图书馆等）的用地需求中分出来的。[41]

专注于证明因果关系的研究一般不会进行这样的成本效益分析，但对于一项干预措施的实施而言，这才是至关重要的。资源不是无限的，实施了一个项目就意味着无法实施另一个项目。[42]在田纳西州的班级规模缩小项目中，项目实施的规模很小，只有那些已经拥有足够的教室、可以开设新班级的学校参与其中。而且这项研究的规模还不足以影响整个地区对教师的需求。

在加州实施这个干预措施之前，为了更好地预测这项政策是否能够取得成效，我们本应该将这项政策的实施背景和面临的限制（比如空间限制）考虑进来，并且判定其他变量会发生什么样的变化（比如这个项目会分走其他项目的资源）。意外的结果会以很多形式出现：一项干预措施可能会有副作用，这就意味着它不仅会导致目标结果，还会导致其他结果，比如一种药物可能会消除患者的头痛症状，但是也会导致患者出现疲惫的症状。然而，这并不会改变系统的性能。但是，人们对自行车共享计划的担忧之一就是它有可能会对健康产生完全负面的影响，比如使用公共自行车的大部分人都是没有经验的骑行者，这就会导致“在城市骑自行车”成为一种很不安全的行为。

这正是加州班级规模缩小计划出问题的地方。新的政策并不能只缩小了班级规模而保持其他所有变量不变。由于这个大项目的实施速度很快，结果导致有些地区的教师质量出现了差异，并且其他项目的实施空间和经费也遭到了缩减。

除了要关注一项干预措施是否会直接实现其目标以外，还要考虑这

项干预措施还可能会导致什么其他的结果。如果我们要预测一个模型，那只需设定班级规模这个变量为真或为假就可以了，但是这个模型无法反映这些情况下将会发生的事情：缩小班级规模是通过经济刺激实现的，而经济刺激的资金又是从其他项目挪出来的，并且新开设的班级没有足够的师资。最终我们还是需要一个更加详细的模型（这个模型既可以是一个我们了解的因果关系模型，也可以是一个我们构建的模拟模型），这个模型不仅要包括一个原因，还要包括实施这个原因的方法，这样的模型能够让我们对比各种缩小班级规模的方法，也就是说，我们可以先针对一些教育水平不高的地区进行实验，然后评估干预措施取得的成效，而不是直接在全州范围内进行推广；也可以先对不同的刺激计划进行测试，等等。当然，并不是所有意外的结果都是负面的。有些意外的结果可能会向我们展示干预措施更多意想不到的好处，从而为某项干预政策的实施提供更多的支持，比如说公共自行车项目降低了空气污染程度，那么这样的意外结果会是一个积极的副作用。

之所以会出现这些副作用，有时是因为我们无法孤立地操纵一件事。我们所实施的不再是“那个干预措施”，而是需要同时改变多个因素。我们不能只是让公共自行车能为人们所用，而是需要像实施自行车共享政策一样来同时实施保护性自行车道政策。这可能是大家都想推广骑行导致的，也可能是作为保证公共自行车项目安全的一个必要条件实施的。

因此，我们可能会在相似的时间段内实施多种政策，这些政策可能也会以无法预测的方式相互作用。比如，一项并不提供头盔的自行车共享项目也许会和一项要求人们使用头盔的法律同时开始实施。如果人们不愿意随身携带他们自己的头盔，那么这项法律可能会减少人们使用共享自行车的次数。由于我们无法立即确定是哪一个干预措施导致了哪些明显的效应，所以多个事件同时改变增加了计划和评估干预措施的难度。然而，如果我们了解了不同的组成部分，就能对它们一一做出解释。[43]

注释

1. 想要阅读这一领域很多研究的文献回顾，参见 Swartz 等（2011）。这个文献回顾后来在内容上又进行了补充和更新，将假设性的食物选择包括了进去（Kiszko 等，2014）。亦见 Krieger 和 Saelens（2013）。
2. Elbel 等（2009）。
3. Carels 等（2007）。
4. Downs 等（2013）。
5. Ellison 等（2014）和 Sonnenberg 等（2013）。
6. Bollinger 等（2011）。
7. 想要了解这方面的例子，参见 Dumanovsky 等（2011）。Dumanovsky 等（2011）在纽约市的相关立法颁布后，考察了一些餐厅菜单的变化。
8. Kearney 和 Levine（2014）。
9. Vandenbroucke（2004）。
10. 正如 Smith 和 Pell（2003）在这篇统计学文章中指出的那样，到目前为止还从来没有出现过一个测试降落伞的 RCT。
11. Hill（1965）。
12. 为什么这些内容不能成为一个清单呢？想要了解更多这方面的信息，参见 Rothman 和 Greenland（2005）和 Phillips 和 Goodman（2004）。
13. 想要了解更多关于 Hill 考虑的因素所起作用的讨论，参见 Höfler（2005）；Ward（2009）。
14. Erickson（1978）。
15. 想要了解关于这一内容的更多讨论，参见 Howick 等（2009）。
16. Schoenfeld 和 Ioannidis（2013）。
17. 想要了解更多关于复制和评估复制的内容，参见 Brandt 等（2014）。
18. 比如说，Hill（1965）就认为这不应该是必要条件，还有更坚定地支持他的看法的人。当然，关于这个条件的批判主要集中在这个条件是否要求各个原因只有一个单一结果这一问题上（Rothman and Greenland，2005）。想要了解关于特异性作用的更为积极的观点，参见 Weiss（2002）。
19. 这个例子来自于 Weiss（2002）。
20. Hanushek（2011）。
21. 参见第 5 章关于这一内容的讨论，并参考 Mill（1843）。
22. Snow（1854）。

23. 第 7 章讨论了各种运行机制。想要了解更多信息，参见 Glennan（1996）和 Machamer 等（2000）。
24. Russo 和 Williamson（2007）。
25. 想要了解更多关于各种类型连贯性的信息，参见 Susser（1991）。
26. 甚至那些使用不同的方法分析同样数据的研究人员对此得出的结论也各不相同（Seok 等，2013；Takao 和 Miyakawa，2014）。
27. 想要回顾这方面的研究，参见 Reiss（2014）。
28. 古德哈特定律本质上是说，一旦我们在政策中使用某个工作指标，这个指标就不再是衡量我们工作业绩的一个准确的指标了。想要了解更多信息，参见 Chrystal 和 Mizen（2003）。
29. 想要了解这方面的例子，参见 Guyatt 等（2008）；Howick 等（2011）。
30. DeMaio（2009）。
31. Goldman 等（2007）。
32. Buck 和 Buehler（2012）。
33. McLean 等（2014）。
34. 想要了解更多关于辅助因素作用的讨论，参见 Cartwright（2012）。
35. 想要回顾在医疗卫生领域将效力转化为效果的各种困难，参见 Glasgow 等（2003）。
36. 想要了解这方面的例子， 参见 Perwien 等（2000）。
37. Blatchford 和 Mortimore（1994）。
38. Bohrnstedt 和 Stecher（2002）。
39. Jepsen 和 Rivkin（2009）。
40. Bohrnstedt 和 Stecher（2002）。
41. Bohrnstedt 和 Stecher（2002）。
42. 比如说，缩小班级规模计划就需要和其他可能带来同样效果的、成本不同的方案进行对比（Normore 和 Ilon，2006）。亦见 Krueger（2003）；Krueger 和 Whitmore（2001）。
43. 比如，Craig 等（2008）就介绍了人们在复杂的医学干预措施上的研究进展，并对这些措施进行了评估，文中的很多指导原则同样适用于很多其他领域。

第 10 章　展望

为什么要研究因果关系？

10.1　人们需要因果关系

自亚里士多德关于因果关系的重要论著问世，已经过去了几千年；自休谟对因果关系提出两个定义，已经过去了几百年；自人们可以通过强大的新型计算机实现因果关系推理自动化，也已经过去了几十年。然而时至今日，因果关系仍然是一个悬而未决的问题。人们一不小心就会推理出一些并不存在的因果关系，而我们的计算程序也不是万无一失的。更糟糕的是，即便我们能够找到一个原因，由于收集和理解信息方面的局限性，我们仍然很难使用这个原因来防止或促使某个结果发生。看完那么多因果分析方法无效的案例和政策制定者彻底弄错因果关系的案例，你可能会想，我们为什么还要如此费力地研究因果关系呢？

在一些小实验中，我们每次只能有规律地改变一个变量来发现一个系统的运行机制。幸运的是，我们现在已经不再受这些小实验的限制了。现在我们拥有了大规模的数据，以数字形式记录了人们的购物习惯、病历以及各种活动。很有可能你随身就携带着手机形式的加速感应器和 GPS 定位器，无论你去哪里它们都会跟着你，而且你的线上活动也在以各种方式被跟踪着。互联网的特性、电子病历的传播以及无处不在的感应器使我

们获得的关于人类的活动数据比历史上任何时候都要多。有了这么多的原始材料，也许事情发生的原因已不再重要。有些人甚至认为，我们通过挖掘这些数据来了解事物之间的相关性就足够了。[1]

有了这么多粒度如此细的数据（比如某个人买书的顺序、人们所走的每一步以及数百万的政治竞选电话的效果），零售商可以针对潜在消费者的情况制作广告，健身公司可以估算出你已经消耗了多少热量，而政治竞选团队则可以找出那些能够被游说的选民。海量数据确实可以让我们的预测更加准确，但如果我们想要知道的只是谁有可能会根据一条广告去买一双鞋子，那么我们也许并不在意这些广告为什么会起作用，也不会在意是否有几个预测弄错了。在这种情况下，就不要去想理论的事了，也不要想着去解释事物发生的原因了，所有的答案都在数据之中。

当然，我们也不是所有时候都会使用“原因”这个词。对这些数据的分析也许可以揭开事物之间的关联性、相关性、关系、纽带和联系，可以揭开事物的发展趋势以及事物发展过程中的风险因素。虽然这些词语的意思相近，但人们常常会将这些发现当成是原因来采取行动。但是，我们使用这些数据主要是为了弄清楚将来会出现什么情况，以便可以改变或者控制将要出现的结果。即使你在工作中并不分析这些数据，也没有兴趣挖掘各种设备（比如健身追踪器）的数据所呈现出的规律，你却无法避免其他人所做的数据分析的结果对你的影响。有一项新政策规定：如果人们佩戴计步器，他们所交的保险费费率就可以降低，你会支持这项政策吗？买药时，你为什么会选择这种药而不选择那种药？在这些情况下，相关性是没有用的。即使我们能够根据相关性成功地预测并干预事件的发展，我们似乎也不可避免地想知道事情为什么会这样发生。从孩子们总喜欢问“为什么”，到成年人总想找到“是谁的错”或者“该怪谁”，我们似乎无一例外地想要知道事情发生的原因。

因果关系绝不像一个多世纪前 Bertrand Russel 说的那样，是“一个已

经过去了的时代的遗迹”。[2]今天，随着我们收集的这些数据集越来越大，因果关系以及对因果关系进行批判性思考的能力比以前任何时候都更加必要。现在对我们来说，知道什么时候能够找到原因以及什么时候找不到原因与会读书写字一样重要。当我们从数百万的测试中提取出一些埋藏在电子碎片中的有意义的信号时，我们很有可能会完全由于偶然因素发现一些似乎具有显著性的结论。因此，我们也越来越需要对任何发现持有怀疑态度。[3]当我们无法通过实验对每一个发现加以验证时，各种统计法可以帮助我们控制虚假发现的数量。但是，如果能知道为什么一个虚假的关系可能会出现在研究结论中，我们就更能将因果关系和相关性区分开来。

人们关于大数据的误解之一就是认为它不过是更多的数据而已——更多的个体、更多的时间点和更多的变量。但是，大数据的收集绝不仅仅是将一个小的数据集扩大而已。要想获得几个电话号码，我们可以查电话号码簿并仔细核对每一个电话号码的真实性。给朋友打电话时，我们清楚地知道手中的电话号码是谁的、这个号码是个人号码还是住宅号码，以及这个号码是手机号码还是固定电话号码。相反，当我们需要数百万的电话号码时，我们根本不可能了解所有号码的每一个使用者，而且我们必须从各种渠道（比如商业数据库和电话号码单）来获得这些电话号码。这些号码可能已经失效或者不准确了，而且我们也无法一个一个地验证这些号码。有些人可能已经搬家了，有些人可能已经将电话号码易主了，还有一些号码可能已经停用了。在大数据库中，出现噪声和错误的概率往往会增加，所以，这种权衡也许并不像人们看到的那么简单。与更小且可控的数据集相比，大数据库中存在更多的数据质量问题、更多潜在的错误来源、更多的偏差以及更多缺失的数据。在海量数据集中，变量更加难以解释，而数据收集的时间表往往也各不相同。大数据并没有让人们不再需要了解事件发生的原因；相反，大数据让事件发生的原因变得更加重要了。

我们不仅需要找到因果关系，还需要对因果所在的领域有深入的了解，这样才有可能知道一个测试是否是成功的，并且能够解释测试得出的结论。在一个研究项目中，我们分析了来自神经科重症监护室的数据，想要了解是什么因素导致中风病人的大脑二次受伤的。在重症监护室，医生通过降低病人的体温来促进病人的康复进程，有些病人的体温甚至下降到了华氏 68 度（约为 20 摄氏度）。这个体温似乎异常的低，但由于这些病人病得很重，所以他们身上的很多指标的数值都是不正常的。如果我们想要知道华氏 68 度是否意味着非常严重的低体温症并且想要对这个数值提出怀疑，就必须预先掌握一些生理学知识。如果我们想要准确地知道为什么会有这么低的体温记录，就要掌握更专业的知识。然而，很多临床医生只要看一眼这个数值就会立刻明白发生了什么。病人的体温是通过插在膀胱里的导尿管测量出来的，所以如果导尿管从膀胱里滑了出来，导尿管测量的温度就成了室内温度，而室内温度恰好在华氏 68 度左右。在了解了这一点之后，事情就很明显了，但只有了解数据以及数据的产生方式的人才能解释为什么会出现这种现象。

如果我们让一个不了解情况的人去数据库里任意发掘，他可能就会错误地发现低体温预示着病人的状况将会得到改善。这是因为导尿管滑出来可能会让护士们对这个病人更加关注，从而迅速发现病人身上可能存在的其他问题。如果我们根据这样一个相关性采取行动的话，就很有可能会采取一些无效的干预措施，从而将病人的体温下降到危险的程度。

除了弄清楚一个数值是否正确以外，想要知道一个变量的含义以及这个变量什么时候会消失，可能也比我们想象的难得多。几乎所有计算法都假设我们已经测量了共同的原因并且手中的变量组合是“正确的”。然而，如果数据不能表明一个变量的真实状态，或者一个共同的结果才是一个原因有没有出现的唯一可靠指标，那么那些假设的条件并不足以让我们搞清楚上述问题。病人的病历中可能会有该病人的诊断结论，之所以会出

现这个诊断结论可能是因为计费的需要，也可能是因为医生怀疑病人有这种症状，还可能是因为病人有得这个病的家族史，或者是因为其他原因（比如复制粘贴错误）。[4]尽管这个值存在于此，但如果它不能准确反映一个病人是否得了某种疾病，那它可能就不能有效排除某个原因所导致的结果，而且这个值的缺失也完全有可能是文件管理失误导致的。如果一个病人确实得了糖尿病，但是这个病却没有恰当地记录在案，那么我们就有可能错误地发现高血糖和胰岛素之间存在相关性。

在某些情况下，我们需要用大量的先验知识来区分在不同的时间表里测量出来的变量（以便测量到所有理论上能够测量到的时间点）和缺乏数据的变量。医院病历数据中的账单代码可以告诉我们病人是由于什么病入院的，有时这些病历还包括病人当前症状的一个清单。如果一个病人的一次就诊记录中出现了哮喘症，但是其他就诊记录中没有出现关于哮喘的记录，那我们又该如何解释这一现象呢？由于哮喘症是慢性病，所以病人只在一次就诊时有哮喘症的可能性不大。但是，病人却有可能只在那一次就诊时治疗了哮喘症（因此只在那一次的记录上出现了哮喘症）。然而，要想知道我们还缺少哪些数据（一名临床医生错误地忽略了疾病清单上的哮喘病）而不是哪些数据是错误的（像流感这样的急性病不会拖很久），我们不仅需要对问题有所了解，还需要对数据产生的方式有所了解。[5]

最好的情况是，那些错误只是随机性噪声，它们会对所有的变量产生相同的影响。但事实上，不同的设备有着不同的错误率，而且人们回答有些问题的准确率可能也会高于另外一些问题。比如，如果我问人们是否吸烟，有些人可能会撒谎，还有一些人可能会将问题理解为他们现在是否在吸烟或者最近是否吸过烟。血压的测量值尤其不可靠，所以我们可能会发现治疗高血压的降压药成了一个人是否患有高血压的指标。我们还会发现这种药物和其他症状之间存在相关性，而不是高血压和其他症状之间存在相关性。我们需要掌握该领域的知识，才能了解这种药物只能表明

哪些人有高血压，以及这种药物本身并不会引起其他疾病。

最后，如果一些大的数据集不是为了研究目的收集的，那么从这些数据集中发现的事物之间的相关性可能不太具有普遍性，这就限制了我们在新环境或未来环境中应用这个结论的能力。2010 年，研究人员对 Facebook 的用户进行了一次测试，想要了解如果他们一登录 Facebook 就会收到关于投票的信息，那么他们在美国国会选举中参与投票的可能性是否会提高，尤其是想要了解如果他们在 Facebook 上看到他们的朋友们都已经投票了，那么他们参与投票的概率是否会提高。[6] 在这次实验中，有 6000 多万用户都收到了社交信息，这些信息向他们展示了他们的一部分朋友已经在 Facebook 上投过票了。还有另外两个小组的用户（每组大约 60 万），其中一个小组只收到了投票信息（比如一个如何找到当地投票点的链接），另一个小组没有收到任何关于选举的信息。通过对比这三个小组的用户投票情况并参考公众投票记录，研究人员估计在 Facebook 上发布的社交信息直接导致投票数增加了约 6 万票（同时又间接导致投票数增加了约 28 万票）。

然而，增加的 6 万多票与 6100 万收到投票信息的人数相比，投票数增加的比例还不足 0.1%。这个绝对数字可能很大，但这次缺乏针对性的实验之所以能够取得这样的成效，完全是由这个巨大的社交网络带来的。如果我们在一个更小的社交网络上复制这一实验，要想让投票人数显著增加，就还需要一个不同并且更直接的方法。事实上，看到好朋友投票的照片似乎要比看到关系比较远的人的投票信息更为有效。但是，进行这样的筛选需要我们了解这些人之间的关系。考虑到这个实验的规模比较小、Facebook 和其他社交网络的用户之间的差别以及各个实验小组规模的差别，我们无法立即确定这个干预措施可以有效应用于其他社交网络或者除美国以外的选举活动。

我们不是要抛弃因果关系，而是要抛弃那种黑匣子思想。不要以为

我们可以从数据源头直接获取一些数据放进黑匣子，然后黑匣子就会吐出一系列无须解释且无须人为干预的原因。因果推理是必要且充满可能性的，却不是完美的。最重要的是，进行因果推理需要掌握因果关系所属领域的专业知识。

10.2 主要原理

研究人员躲在他们各自所属学科的筒仓里，就发现和使用原因的最好的方法争论不休。这很容易让人们形成这样印象，认为有很多互不相干的领域在独立地对问题的一个个微小部分起作用。这些研究人员没有达成任何明显的共识，而且每种方法的众多局限性导致整个研究似乎已经陷入绝境。即使我们真的需要原因，最后可能也无法找到它们。

确实，因果关系问题尚未解决，而且也没有能够适用于所有因果关系问题的理论。我们无法给原因下一个适用于所有案例的定义，也没有一个可以从所有类型的数据中找到原因的方法。研究人员也许会因为这个领域中还要很多未知空间而对因果研究充满热情，但如果你不是一名研究人员，你又能从这个领域学到什么呢？

虽然我们并未掌握关于因果关系的所有信息，但也确实掌握了一些信息。更重要且更激励人心的是，我们对于因果关系的认识已经随着时间的变化更为深入了。我们之所以能够加深对因果关系的认识，一方面是因为我们有了更好的数据和更强大的计算能力，另一方面是因为各个领域的不断重叠和跨学科研究不断发展。

10.2.1 因果关系和相关性不是同义词

本书最大的启示之一在于，要想找到事情发生的原因绝不是一件容易的事。很多时候我们认为自己已经找到了原因，但实际上我们找到的只

是事物之间的相关性而已，而且有时连这些相关性都是假的。这可能是概念混乱造成的（考察的变量不正确，结果在由一个共同原因导致的多个结果之间找到了一个虚假的关系），也可能是我们在寻找和评估信息的方式存在偏差导致的（证实性偏差意味着我们只会去寻找那些正面的例子），还可能是我们考察的很多其他因素导致的。

有很多方法可以让我们在没有因果关系的情况下也能发现事物之间的相关性（反之亦然）。知道这些方法很重要，因为它能帮助我们批判性地评估我们的发现和假设，还可以防止我们去采取一些无效的干预措施。假设某人发现他跑步的距离和他的精力水平之间存在相关性。令人感到意外的是，他跑步的距离越长，他的精力似乎就越旺盛。但是，如果只有在空闲时间比较多且能够睡懒觉的日子里，他的跑步距离才会变长，那么他真正会发现的结论应该是睡觉的时间越长，他的精力就越旺盛。在这种情况下，如果他预测跑完一场马拉松比赛后，他的能量一定会激增，那么这样的预测一定会失败。这意味着对他来说，要想感到精力充沛，最好的策略就是多睡觉而不是多跑步。

不论数据的大小如何，我们都必须对我们的发现进行质疑，要多问问“为什么”。比如，通过使用人们的搜索词和流感病例之间的相关性，谷歌预测出流感趋势的时间比疾控中心还要早。[7] 但是，这种方法只有在这样的前提条件下才会有效：人们主要在有流感症状时才会使用谷歌搜索引擎检索这些搜索词；人们进行搜索不是因为他们担心流感暴发，也不是因为家里有人出现了流感症状，更不是因为他们听说了谷歌的这个研究项目。事实上，随着时间的变化，谷歌预测流感趋势的表现越来越差。2011 年，谷歌预测的流感严重程度要比实际观察到的流感严重程度高得多，而且在之后的很长一段时间里，它预测的数据都比实际数据要高。[8] 如果我们不清楚为什么某件事情具有预测作用的话，那我们的预测就不可避免地会出现意料之外的失败。

10.2.2 针对偏差的批判性思考

虽然有很多容易出错的地方，但如果能够找出这些容易出错的地方并因此提高警惕，我们就能够设计出更好的方法和更加有效的干预措施，还能避免推理出错误的结论。我们之所以要用一整章的内容来讨论因果关系的心理学问题，是因为如果了解了我们在哪些地方擅长寻找原因，就能够设计出更好的方法来让这一过程自动化，同时也是因为如果能找到我们思维中容易出错的地方，就能更好地处理这些薄弱区域。这可能意味着我们在避免认知偏差时要提高警觉，[9] 意味着我们要设计出一些可以更好地处理选择性偏差的计算程序，[10] 或者意味着我们要将数据清理和分析任务交给不同的、对研究假设一无所知的人来做，以避免无意中犯下证实性偏差这样的错误。[11]

心理学为我们理解一些长期存在的哲学问题（比如道德评价和因果判断之间的关系）提供了一些启示，同时还建议我们应该更加关注外部有效性以及我们对推理方法和解释方法的评价。

在很多情况下，我们需要从不同来源收集的数据要比一开始计划的多得多。心理学领域有一个重要发现：人们可能不仅在引起某个事件的原因上存在分歧，而且在看待引起同一个事件的不同原因的相对显著性上也存在分歧。这些分歧可能来自于文化上的差异，在我们设计用来寻找原因的各种方法时，有必要意识到这种文化差异。在哲学上，人们经常通过分析来评估一些案例，以便了解某个理论是否给出了我们想要的答案，这就表明一个人的直觉并不一定具有普遍性。

加拿大的某位教授认为的导致某人考试作弊的原因也许和印度某位农民的想法不一致。甚至在 Michotte 的这种简单的因果关系感知研究中，也不是所有参与者都会以同样的方式来感知一些场景。实体因果关系往往更加复杂，很多不同的答案都有可能是正确的，而且或多或少在不同的案

例中都有所关联。交通事故可能是多种原因共同导致的，比如汽车生产商的失职、驾驶员注意力不集中以及恶劣的天气情况。但是，法律案件关注的重点和其他案件关注的重点是不一样的。这些解释上的差异还会影响陪审团的决定，并最终影响陪审团的甄选过程。

实验哲学领域的研究曾试图明确陪审团做的这些判断到底有多大的差别，并试图找到是哪些因素导致陪审团形成了不同的观点并改变了评估案件的方式。尽管目前还没有一个完美的理论能够解释人们是如何划分责任或确定实体原因的，但使用认知心理学的实验法来解决哲学问题可以帮助我们在评估这些方法的过程中摆脱对个体直觉的依赖。

为了验证这些方法的有效性，我们需要准确客观的真相（导致某个事件发生的真实原因），以此来衡量使用各种方法得到的结论。但是，如果解释是主观的，真相也是因人而异的，那么我们就需要重新评估我们的验证方案了。比如说，如果我们对众包工人（比如亚马逊土耳其机器人）进行民意测验，或者对某个社交网络的使用者进行问卷调查，那么我们就应该更关注结论中的文化偏差，并且在多个文化背景下复制这项调查，以保证参与者在人口统计学上的多样性。

10.2.3　时间的重要性

1948 年美国总统大选当晚，《芝加哥论坛报》的大字标题印刷错误，写成了“杜威打败杜鲁门”。[12] 这件事在当时非常有名。这份报纸必须在选举结果确定之前付印，而当时 Gallup、Roper 和 Crossley 所做的民意调查都预测杜威会取得决定性的胜利。之所以会出现这样的情况，一方面是由于取样方法不当，导致共和党人在样本中所占的比例过高，另一方面是由于这些机构停止进行民意调查的时间过早，有些机构甚至在 9 月份（选举正式举行之前的两个月）就停止民意调查了。[13] 他们假设人们是否会投票以及打算把票投给谁这一结果在最后的几个月内不会发生改变。而且这

些民意测验的结果可能也影响到了选举。“杜威似乎很明显会成为最后的赢家”的念头可能会让他的支持者们过于自信，从而大大降低了在选举当日去投票的可能性。相反，杜鲁门的支持者们可能会由于民意测验的情况不好而受到激励，导致他们去参加投票的可能性大大增加。

同理，人们在计算疾病风险时可能也会因为使用了历史数据而高估了疾病在当前人口中暴发的风险。我们必须搞清楚数据和因果关系有没有可能随着时间的变化而发生改变，以及它们在我们研究的那个时刻是否仍然适用。

不管是寻找物理事件中的因果关系（在缺乏机械知识的情况下，时间上的滞后会导致人们认为事物之间可能存在因果关系的可能性下降），还是评估干预政策（在评估干预政策时，需要根据时间因素来评估风险并判断原因的效力），我们都不能忽略事件中的时间因素。由于我们希望结果紧随原因之后出现，所以时间对于我们感知事物之间的因果关系极为重要。如果我们对原因导致结果的过程（比如吸烟需要很久才能导致癌症）有所了解，可能就会理解原因和结果之间的时间间隔。但是“原因出现在结果之前”这一思想对于我们考察过的很多哲学理论来说都至关重要，并且这一思想也得到了心理学实验的支持。

10.2.4 并不是所有实验研究都比观察性研究好

是采用观察性研究还是实验研究，这个问题是一个错误的两分法。实际上，我们不可能在每一种情况下都能进行实验研究（有人愿意在跳伞过程中作为对照组去研究降落伞能否防止死亡事件的发生吗），而且也不总是需要这样做（物理学和工程学再加上一些模拟，这就可以代替一个关于降落伞的 RCT）。更重要的是，有很多方法可能会导致随机试验的失败，而且在有些情况下，我们也可以通过观察来了解事情发生的原因。

由于对医学研究进展缓慢而感到沮丧，一帮患有肌萎缩侧索硬化症

（ALS）的病人自己设计了一项研究，来测试一种实验疗法能否减慢他们疾病的恶化速度。[14]对于这种由病人领导的研究来说，由于这些病人对他们的健康问题积极性很高，所以他们所面临的难题就是如何设置对照组。实际上，这项试验研究使用了参与者和来自社交网站 PatientsLikeMe（像我一样的病人）的其他病人分享的大量数据。在医生们的支持下，实验组在他们的治疗方案中加入了锂，并且自我跟踪研究了 12 个月，详细记录了他们的状况。

由于这个试验不是盲法试验，也不是随机试验，所以这项研究很容易出现很多偏差。为了解决这个问题，每一个病人都与很多个没有服用锂的病人搭配在一起进行对照，在试验开始时，这些作为对照的病人与参与实验的病人病情相似。在接受锂治疗后，通过与这些病人进行对照，参与实验的病人可以看出他们和对照病人在病情上是否有差异。结果没有发现任何差异。这一负面结论在随后进行的多次随机试验中得到了证实。[15]由于很多因素都可能会出现偏差，从而导致结果对这种药物有利，所以从某种意义上来说，在这群病人中得出的负面结论比在一次 RCT 中得出的结论更为有力。病人进行的不是盲法试验，治疗效果也是由病人自己报告的，而且由于他们希望这个药物有效，所以那些认知偏差完全有可能导致他们以不同的方式来评估他们的健康状况。在很多情况下，将实验数据和观察数据仔细地结合在一起可以解决彼此的局限性问题。此外，当这两个数据得出的结论一致时，会增强人们对这两种方法的信心。

10.3　一个百宝箱

如果你只有一把锤子，那么每一个问题都像是一根钉子。我们之所以要过于详细地讨论每一种方法的缺点，并不是为了让人们觉得哪一种方法都不行，而是想要说明没有哪一种方法是万能的。概率模型不是因果推

理的唯一方法，反事实推理法也不是解释事件的唯一方法。很多方法都在以人们意想不到的方式被应用于不同的学科当中。格兰杰因果理论一开始是用于经济时间序列中的，但是现在已经被用来分析神经元放电活动记录了。[16]人们开发贝叶斯网络是为了表示概率性关系，但它现在已经被用来模拟因果推理背后的心理过程了。[17]没有任何方法或模型是放之四海皆准的，你可能需要超越自己的研究领域来寻找问题的解决方案。

如果这个解决方案有标准答案，那么这个答案就是我们需要利用多种方法。每一种方法适用于一种不同的情况，所以，如果你只有一种运用自如的工具，那么你会由于这个工具的局限性而挫败不已。只要付出心血，大部分情况下都能将这个工具改变得可以适用于各种情况。但是，我们并不是要你用胶带和铁片对一个锤子进行改造，然后用它来翻动煎饼。如果你知道一个叫铲子的东西，那么你就可以省去很多麻烦。

近几年来，人们越来越意识到我们需要的是一组能够互补的方法，而不是一种能够解决所有问题的方法。[18]比如，Illari 和 Russo 最近提出了关于因果关系的拼贴观。就像一片瓷砖在一幅图画中的作用不能仅通过它本身来理解，我们需要使用的方法取决于问题的背景、意义、眼前的问题和我们的目的。

这是因果关系多元化趋势的一部分，还有很多事物都可以被多元化。我们可以让原因的定义多元化，[19]让支持原因的证据多元化，还可以让收集证据的方法多元化。[20]从实际出发，我们通常不太关心因果关系在形而上学层面的研究，或者说不太关心“原因究竟是什么”的研究。但是，前面最后两点之间的差别还是值得注意的。人们可能会认为能够通过多种特征来将因果关系和相关性区分开来，比如通过概率法、干预法和机械法都能对原因有所了解。但是，在这些方法中间，即使你认为干预法是唯一可以支持某个因果假设的方法，用来收集干预法所需证据的方法也有很多个（只要想想第 7 章介绍的各种实验法就能明白了）。同理，因果关系显著性

的测量方法也有很多，这些方法强调的特征也各不相同。

针对机器学习过程中的一些问题（比如优化问题），有一组法则叫作“没有免费的午餐”。[21] 就是说如果一个方法是针对某一种问题设计出来的，那么这种方法就很难解决其他类型的问题，没有任何方法能够完美适用于所有的测试。这就意味着我们的方法不可能完美解决所有问题，如果一种方法将一个问题解决得很好，那么它在另一个问题上就不可能不需要调整。这看起来可能很麻烦，因为如果我们面对的是一个新问题，我们就不知道该用哪一种方法了。

但是，我们也并不总是在一无所知的情况下挑战一个新问题。如果我们对手头上的这个问题已经有所了解，也知道我们愿意做出什么样的让步（比如接受更多的假阴性以便减少假阳性结果），那么我们并不需要一个完美适用于所有情况的方法，只需知道如何针对正在解决的某个问题来选择一个更好的解决方案就可以了。

比如，如果我们想考察在某个城市的餐厅内公布食物热量值是否真的会降低人们所消费食物中的热量值，这就是一个关于实体因果关系的问题，它更适合用反事实推理法来解决，而不是格兰杰的因果关系理论。相反，假设我们现在有来自计步器和联网电子秤的数据，也知道人们摄入的热量值。在这种情况下，如果我们想根据人们的锻炼情况和饮食习惯来预测人们的体重，那么我们要解决的问题就是一个完全不同的问题了，解决方法也会与上面的案例大不相同。对于这个问题来说，贝叶斯网络可能是一个很好的选择。因为贝叶斯网络更擅长根据网络中其他变量的值来预测某个变量可能会出现的值。然而，如果我们想知道在高强度的运动之后需要多久血糖才会升高，选择贝叶斯网络就不是什么好主意了，我们应该选择一个能够让我们从数据中发现某种关系的时间性的方法。

最重要的是，我们对因果关系还有很多不了解的地方。如果只局限于改造已有的方法，可能就会错过很多重要的发现。

10.4 知识的重要性

随着不断设计出更好的方法来寻找原因并预测未来要发生的事件，我们可能想让更多的发现过程能够自动化，并慢慢将人从这一过程中剥离出去。毕竟人是有偏差、不理性且无法预测的，而计算机程序在每次接受同样的指令后，都会以完全相同的方式忠实地执行。但目前来讲，寻找原因的每一步都离不开人的知识和判断：决定收集什么样的信息、准备数据、选择数据分析方法、解释研究结果以及决定如何根据研究结果采取行动。

我们曾经想要找到这么一个黑匣子，让它能在没有人为输入的情况下，百发百中地将“原始”数据顺利地转化成原因。通过前面的研究，我们已经知道这种想法错误的原因是什么。但是，用这种没有人为判断的方式来使用原因也是错误的。如果一个公司为一种你不感兴趣的产品做广告，或者一个网站推荐了一部你不喜欢的电影，这些错误的成本很低。但在其他很多情况下（比如 Sally Clark 冤案），误用因果关系可能会带来严重的后果。我们可能会过度信任一个推理出来的结论，或者某种计算方法可能会导致我们过度依赖一般性的常识而不考虑具体情形中的具体细节。

医生说你的血压太高，需要采取一些措施。这时，如果他盲目地按照一套指导原则来给你开处方，你肯定不愿意；相反，你希望他在开药时能考虑一下你当前可能正在服用的其他药物（这些药物有可能会与降压药相互作用），并且希望他能考虑你自己的治疗重点和目标。根据治疗高血压的一般性指导原则，医生最后开出的处方可能不是最好的治疗方案，但却有可能是对于你个人来说最好的治疗方案。因为尽管高血压可能会导致严重的健康问题，但是降低血压并不是你的唯一目标，降低血压必须和其他目标结合在一起来考虑。你可能正在服用一些药物，而这些药物会与指导原则推荐的降压药产生药物反应。而且在降压药的服药次数上，你可能更容易遵守每日一次的服药要求，而不是每日多次的服药要求。[22]也有

可能由于医疗保险的要求，你不得不遵守一些用药方面的限制。前面已经说过，我们不能仅根据一个已知类型层面的关系去推理一个实体层面的原因；同理，我们也不应该只使用类型层面的信息去做关于实体案例的决策。

找到原因后，当我们在考虑如何使用原因以及是否应该使用原因时，我们需要考虑的不仅仅是关系的有效性问题。

在美国，至少有 20 个州已经采用了某种形式的基于证据的刑事量刑制度。这一制度通过计算犯人将来再次犯罪的风险来指导量刑。[23] 医学领域通过标准化进程已经取得了很大的进展，能够保证根据证据而不是直觉去为病人提供标准的、优质的医疗服务。基于证据的刑事量刑制度也正试图为人们提供一个更加公正的方式来确定犯人对社会的威胁，并减少由于法官的辨别力或判断力不同而可能导致的偏差。我们很难去反对这一量刑制度的原则和目标。

然而，这些风险计算器考虑了很多与犯人的犯罪记录无关的其他特征（比如经济状况和就业情况），并且包括了一些个人无法控制的因素（比如性别）。这就意味着如果两个人犯了同样的罪，如果其中一个人生活在一个犯罪率比较低的社区，或者如果他有一份稳定的工作，那么人们可能会认为他再次犯罪的可能性比较低。尽管这些因素中并没有直接包含种族因素，但是其中的很多因素都和种族因素相关。这种方法和犯人有没有犯罪记录没有任何关系，和这些因素是否与犯罪行为相关也没有任何关系。相反，这种方法更像是保险公司使用寿险精算表来给保险产品定价一样。一个人的预期寿命其实是不可知的，所以这些表格根据客户所属群体（比如年龄和性别群体）的预期寿命来计算具体客户的预期寿命。

先不要管不同的刑期是否真的会让犯人再次犯罪的可能性降低，也不要管我们测量的（关于各种特征的人有多少是再次犯案的犯人）数据是

否正确，[24]先问问应不应该用这个信息来决定犯人刑期的长短。

一个原因可以用来做出准确的预测或者可以用来指导人们的决策行为并不意味着这个原因应该被用来做这样的事。因果推理方法只能告诉我们某些群体再次犯罪的概率是不是高一些，但不能告诉我们一个公正的社会是否应该使用这样的群体特征来更加严厉地惩罚某些罪犯。挖掘大型数据集来寻找事物相关性的风险之一就是我们不知道事情为什么会发生，虽然因果推理会通过客观性给人一种公正的表象，但也可能会被用来支持一些不公正的、带有歧视性的行为。负责任地使用原因意味着我们不仅要评估我们的发现在统计学和方法论上的合理性，还要评估这些发现的后果和道德基础。

我们需要将人类能够在深思熟虑的基础之上进行判断的优势和计算机能够以一种人类无法企及的方式对海量数据进行挖掘的优势结合在一起，而不是让所有事情完全自动化。无论什么时候，只要我们面对一种可能的因果关系，都必须要找到能够支持这种因果关系的证据，同时还要像对待犯人一样审问它：我们的证据仅仅是间接证据吗（就像相关性一样），还是背后有什么动机（为什么某个原因会导致某个结果的一个机械性解释）？有没有什么减轻责任的因素（比如一个共同原因或者数据中的一些偏差）？随着与我们的发现有关的成本和风险的上升，证据所承担的压力也必须加大。当我们无法非常自信地找到原因时，必须勇于将这种不确定性说出来，告诉人们我们确实不知道原因是什么——然后再接着找。

注释

1. Chris Anderson 早在 2008 年就在《连线》杂志的网站上提出了观点——拍字节让我们可以说，“有相关性就够了”。（Anderson，2008）。
2. Russell（1912）。
3. 参见第 3 章关于多重对比的讨论。

4. 想要了解更多关于影响诊断代码准确性因素的信息，参见 O'Malley 等（2005）。
5. 想要了解更多根据病历记录规律来区分慢性病和急性病的信息，参见 Perotte 和 Hripcsak（2013）。
6. Bond 等（2012）。
7. Ginsberg 等（2009）。
8. Lazer 等（2014）。
9. 注意，意识到偏差的存在并不意味着我们就能完全避免这些偏差。想要从非技术层面了解这种偏差在决策过程中的表现，参见 Kahneman 等（2011）。
10. 想要了解一些例子，参见 Bareinboim 和 Pearl（2012）；Robins 等（2000）；Spirtes 等（1995）。
11. 想要了解更多方法论方面考虑的问题，参见 Young 和 Karr（2011）。
12. Henning（1948）。
13. Mitofsky（1998）；Sudman 和 Blair（1999）。
14. Wicks 等（2011）。
15. 想要了解关于 ALS 治疗方案研究的更为广泛的讨论，或想要了解更多关于锂治疗方案的不同研究，参见 Mitsumoto 等（2014）。
16. 事实上，在为数不多的、包含多变量格兰杰因果关系的软件包中，就有一个软件是神经学家开发出来的（Barnett 和 Seth，2014）。
17. 想要回顾这方面的研究，参见 Holyoak 和 Cheng（2011）。
18. 想要回顾这方面的研究，参见 Godfrey-Smith（2010）。
19. 这就叫形而上学的多元论（Psillos，2010）。
20. Russo（2006）。
21. 想要一个简练的解释，参见 Ho 和 Pepyne（2002）；想要一个更加深入的解释，参见 Wolpert 和 Macready（1997）。
22. 有很多研究考察了给药方案和人们遵守这一方案的程度之间的联系。想要回顾这方面的研究，参见 Claxton 等（2001）。
23. 想要回顾这方面的内容，参见 Slobogin（2012）；想要了解关于问题和伦理的讨论，参见 Sidhu（2015）；Starr（2014）。
24. 我们试图通过对逮捕人数和报案次数进行对比，来证实这些方法是有效的，但这仍然不能告诉我们实际发生了多少起案件，只能告诉我们有多少人被捕了。即使犯罪活动水平是一样的，有些社区的逮捕率也可能会高于其他社区。

致谢

如果没有资助机构对我关于因果关系研究工作的支持，就不可能有这本书。我的研究工作和创作过程得到了美国国立卫生研究院国家医学图书馆（资助编号：RoIMoII826）和美国国家科学基金会（项目编号：I347II9）的资助。本书提到的任何观点、发现以及结论或建议都是作者的个人意见，并不一定反映美国国立卫生研究院和美国国家科学基金会的观点。

谨以此书献给我的母亲，她才是此书得以成书的真正原因。

参考文献

Afari, N. and Buchwald, D. (2003). Chronic Fatigue Syndrome: A Review. *American Journal of Psychiatry*, 160(2):221–236.

Ahn, W.-K. and Bailenson, J. (1996). Causal Attribution as a Search for Underlying Mechanisms:An Explanation of the Conjunction Fallacy and the Discounting Principle. *Cognitive Psychology*, 31(1):82–123.

Ahn, W.-K. and Kalish, C. W. (2000). The role of mechanism beliefs in causal reasoning. In F. C. Keil and R. A. Wilson (eds.), *Explanation and cognition*, pp. 199–225. The MIT Press, Cambridge, MA.

Ahn, W.-K., Kalish, C. W., Medin, D. L., and Gelman, S. A. (1995). The role of covariation versus mechanism information in causal attribution. *Cognition*, 54(3): 299–352.

Alberts, B. (2011). Retraction of Lombardi et al. *Science*, 334(6063):1636–1636.

Alexander, J. (2012). *Experimental philosophy: An introduction*. Polity, Cambridge, UK.

Alicke, M. D., Rose, D., and Bloom, D. (2011). Causation, Norm Violation, and Culpable Control. *The Journal of Philosophy*, 108(12):670–696.

Alter, H. J., Mikovits, J. A., Switzer, W. M., Ruscetti, F. W., Lo, S.-C., Klimas, N., Komaroff, A. L., Montoya, J. G., Bateman, L., Levine, S., Peterson, D., Levin, B., Hanson, M. R., Genfi, A., Bhat, M., Zheng, H., Wang, R., Li, B., Hung, G.-C., Lee, L. L., Sameroff, S., Heneine, W., Coffin, J., Hornig, M., and Lipkin, W. I. (2012). A Multicenter Blinded Analysis Indicates No Association between Chronic Fatigue Syndrome/Myalgic Encephalomyelitis and either Xenotropic Murine Leukemia Virus-Related Virus or Polytropic Murine Leukemia Virus. *mBio*, 3(5):e00266–12.

Andersen, H. (2013). When to Expect Violations of Causal Faithfulness and Why It Matters. *Philosophy of Science*, 80(5):672–683.

Anderson, C. (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete.

Appelbaum, B. (2011). Employment Data May Be the Key to the President's Job. *The New York Times*, p. A1.

Aristotle (1924). *Metaphysics*. Oxford University Press, Oxford. Edited by W. D. Ross.

——— (1936). *Physics*. Oxford University Press, Oxford. Edited by W. D. Ross.

Badler, J., Lefèvre, P., and Missal, M. (2010). Causality Attribution Biases Oculomotor Responses. *The Journal of Neuroscience*, 30(31):10517–10525.

Badler, J. B., Lefèvre, P., and Missal, M. (2012). Divergence between oculomotor and perceptual causality. *Journal of Vision*, 12(5):3.

Baird, S., Ferreira, F. H. G., Özler, B., and Woolcock, M. (2013). Relative Effectiveness of Conditional and Unconditional Cash Transfers for Schooling Outcomes in Developing Countries: A Systematic Review. *Campbell Systematic Reviews*, 9(8).

Baker, S. G. and Kramer, B. S. (2001). Good for Women, Good for Men, Bad for People: Simpson's Paradox and the Importance of Sex-Specific Analysis in Observational Studies. *Journal of Women's Health & Gender-Based Medicine*, 10(9): 867–872.

Bareinboim, E. and Pearl, J. (2012). Controlling selection bias in causal inference. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*.

Barnett, L. and Seth, A. K. (2014). The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68.

Beasley, N. A. (1968). The extent of individual differences in the perception of causality. *Canadian Journal of Psychology*, 22(5):399–407.

Bechlivanidis, C. and Lagnado, D. A. (2013). Does the "Why" Tell Us the "When"? *Psychological Science*, 24(8):1563–1572.

Beecher, H. K. (1955). The Powerful Placebo. *Journal of the American Medical Association*, 159(17):1602–1606.

Begley, C. G. and Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533.

Bennett, C. M., Baird, A. A., Miller, M. B., and Wolford, G. L. (2011). Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, 1:1–5.

Bhatt, A. (2010). Evolution of Clinical Research: A History Before and Beyond James Lind. *Perspectives in Clinical Research*, 1(1):6–10.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Blackwell, B., Bloomfield, S. S., and Buncher, C. R. (1972). Demonstration to medical students of placebo responses and non-drug factors. *The Lancet*, 299(7763): 1279–1282.

Blatchford, P. and Mortimore, P. (1994). The Issue of Class Size for Young Children in Schools: What can we learn from research? *Oxford Review of Education*, 20(4): 411–428.

Bohrnstedt, G. W. and Stecher, B. M. (eds.) (2002). *What We Have Learned about Class Size Reduction in California*. American Institutes for Research, Palo Alto, CA.

Bollinger, B., Leslie, P., and Sorensen, A. (2011).Calorie Posting in Chain Restaurants. *American Economic Journal: Economic Policy*, 3(1):91–128.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D., Marlow, C., Settle, J. E., and Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298.

Born, M. and Einstein, A. (1971). *The Born Einstein Letters: Correspondence between Albert Einstein and Max and Hedwig Born from 1916 to 1955 with commentaries by Max Born*. Macmillan Press, Basingstroke, UK. Translated by Irene Born.

Boyd, C. M., Darer, J., Boult, C., Fried, L. P., Boult, L., and Wu, A. W. (2005). Clinical Practice Guidelines and Quality of Care for Older Patients With Multiple Comorbid Diseases: Implications for Pay for Performance. *JAMA*, 294(6):716–724.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., and Van't Veer, A. (2014). The Replication Recipe: What makes for a

convincing replication? *Journal of Experimental Social Psychology*, 50:217–224.

Broadie, S. (2009). The Ancient Greeks. In H. Beebee, C. Hitchcock, and P. Menzies (eds.), *The Oxford Handbook of Causation*, pp. 21–39. Oxford University Press, Oxford; New York.

Buchanan, M. (2007). Statistics: Conviction by numbers. *Nature*, 445:254–255.

Buck, D. and Buehler, R. (2012). Bike Lanes and Other Determinants of Capital Bikeshare Trips. In *91st Transportation Research Board Annual Meeting*.

Buehner, M. J. and May, J. (2003). Rethinking temporal contiguity and the judgement of causality: Effects of prior knowledge, experience, and reinforcement procedure. *The Quarterly Journal of Experimental Psychology, Section A*, 56(5):865–890.

——— (2004). Abolishing the effect of reinforcement delay on human causal learning. *The Quarterly Journal of Experimental Psychology, Section B*, 57(2):179–191.

Buehner, M. J. and McGregor, S. (2006). Temporal delays can facilitate causal attribution: Towards a general timeframe bias in causal induction. *Thinking & Reasoning*, 12(4):353– 378.

Campbell, M. K., Elbourne, D. R., and Altman, D. G. (2004). CONSORT statement: Extension to cluster randomised trials. *BMJ*, 328:702–708.

Caporael, L. R. (1976). Ergotism: The Satan Loosed in Salem. *Science*, 192(4234): 21–26.

Carels, R. A., Konrad, K., and Harper, J. (2007). Individual differences in food perceptions and calorie estimation: An examination of dieting status, weight, and gender. *Appetite*, 49(2):450–458.

Carey, B. (2012). Father's Age Is Linked to Risk of Autism and Schizophrenia. *The New York Times*, p. A1.

——— (2013). Sleep Therapy Seen as an Aid for Depression. *The New York Times*, p. A1.

Carpenter, C. E. (1932). Workable Rules for Determining Proximate Cause. *California Law Review*, 20(3):229–259.

Cartwright, N. (1999). Causal Diversity and the Markov Condition. *Synthese*, 121(1-2):3–27.

——— (2001). What Is Wrong with Bayes Nets? *The Monist*, 84(2):242–264.

——— (2002). Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward. *British Journal for the Philosophy of Science*, 53(3):411–453.

——— (2004). Causation: One Word, Many Things. *Philosophy of Science*, 71(5): 805–819.

——— (2012). Presidential Address: Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps. *Philosophy of Science*, 79(5):973–989.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., and Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, 1(1):4–17.

Charney, E. and English, W. (2012). Candidate Genes and Political Behavior. *American Political Science Review*, 106(1):1–34.

Charniak, E. (1991). Bayesian Networks without Tears. *AI magazine*, 12(4):50–63.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological review*, 104(2):367–405.

Cheng, P. W. and Novick, L. R. (1990). A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology*, 58(4):545–567.

——— (1992). Covariation in natural causal induction. *Psychological Review*, 99(2):365–382.

Cherry, W. H. and Oldford, R. W. (2003). Picturing Probability: The poverty of Venn diagrams, the richness of Eikosograms. Unpublished manuscript.

Choi, I., Dalal, R., Chu, K.-P., and Park, H. (2003). Culture and Judgement of Causal Relevance. *Journal of Personality and Social Psychology*, 84(1):46–59.

Choi, I., Nisbett, R. E., and Norenzayan, A. (1999). Causal Attribution Across Cultures: Variation and Universality. *Psychological Bulletin*, 125(1):47–63.

Chrystal, K. A. and Mizen, P. (2003). Goodhart's Law: Its origins, meaning and implications for monetary policy. In P. Mizen (ed.), *Central Banking, Monetary Theory and Practice: Essays in Honour of Charles Goodhart*, volume 1, pp. 221–243. Edward Elgar Publishing, Northampton, MA.

Chua, H. F., Boland, J. E., and Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences*, 102(35): 12629–12633.

Claxton, A. J., Cramer, J., and Pierce, C. (2001). A systematic review of the associations between dose regimens and medication compliance. *Clinical Therapeutics*, 23(8): 1296–1310.

Cohen, J. (2011). Chronic fatigue syndrome researcher fired amidst new controversy. *Science*.

Cohen, L. B., Rundell, L. J., Spellman, B. A., and Cashon, C. H. (1999). Infants' perception of causal chains. *Psychological Science*, 10(5):412–418.

Collins, H. and Pinch, T. (2008). *Dr. Golem: How to Think about Medicine*. University of Chicago Press, Chicago.

Conley, R. H. and Conley, J. M. (2009). Stories from the Jury Room: How Jurors Use Narrative to Process Evidence. *Studies in Law, Politics, and Society*, 49(2):25–56.

Cook, N. R. and Ridker, P. M. (2014). Response to Comment on the Reports of Overestimation of ASCVD Risk Using the 2013 AHA/ACC Risk Equation. *Circulation*, 129(2):268–269.

Cooke, P. (2009). Clarifications and corrections to 'On the attribution of probabilities to the causes of disease' by Peter Cooke and Arianna Cowling (Law, Probability and Risk (2005), 4, 251–256). *Law, Probability & Risk*, 8:67–68.

Cooke, P. and Cowling, A. (2006). On the attribution of probabilities to the causes of disease. *Law, Probability & Risk*, 4(4):251–256.

Cooper, G. F. (1999). An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks. In C. Glymour and G. F. Cooper (eds.), *Computation, Causation, and Discovery*, pp. 3–62. AAAI Press and MIT Press, Cambridge, MA.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, 9(4):309–347.

Corrao, G., Rubbiati, L., Bagnardi, V., Zambon, A., and Poikolainen, K. (2000). Alcohol and coronary heart disease: A meta-analysis. *Addiction*, 95(10):1505–1523.

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., and Petticrew, M. (2008). Developing and evaluating complex interventions: The new Medical Research Council guidance. *BMJ*, 337:a1655.

Crofton, J. (2006). The MRC randomized trial of streptomycin and its legacy: A view from the clinical front line. *Journal of the Royal Society of Medicine*, 99(10): 531–534.

Cushing, J. T. (1998). *Philosophical Concepts in Physics*. Cambridge University Press, Cambridge.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, 108(2):353–380.

Dalakas, M. C. (1995). Post-Polio Syndrome As an Evolved Clinical Entity. *Annals of the New York Academy of Sciences*, 753:68–80.

Damisch, L., Stoberock, B., and Mussweiler, T. (2010). Keep Your Fingers Crossed! How Superstition Improves Performance. *Psychological Science*, 21(7):1014–1020.

Danks, D. (2005). The Supposed Competition Between Theories of Human Causal Inference. *Philosophical Psychology*, 18(2):259–272.

Dash, D., Voortman, M., and De Jongh, M. (2013). Sequences of mechanisms for causal reasoning in artificial intelligence. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*.

David, L., Seinfeld, J., and Goldman, M. (writers) and Cherones, T. (director). (1991). The stranded [Television series episode]. In David, L. (producer), *Seinfeld*. CBS, Los Angeles.

DeMaio, P. (2009). Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of Public Transportation*, 12(4):41–56.

Devine, D. J., Clayton, L. D., Dunford, B. B., Seying, R., and Pryce, J. (2001). Jury decision making: 45 years of empirical research on deliberating groups. *Psychology, Public Policy, and Law*, 7(3):622–727.

Diamond, S. S. and Rose, M. R. (2005). Real Juries. *Annual Review of Law and Social Science*, 1:255–284.

Diamond, S. S., Vidmar, N., Rose, M., Ellis, L., and Murphy, B. (2003). Juror Discussions during Civil Trials: Studying an Arizona Innovation. *Arizona Law Review*, 45:1–83.

Dickey, D. A. and Fuller, W. A. (1981). Likelihood Ratio Statistics for Autoregressive Time Series with a Unit Root. *Econometrica*, 49(4):1057–1072.

Downs, J. S., Wisdom, J., Wansink, B., and Loewenstein, G. (2013). Supplementing Menu Labeling With Calorie Recommendations to Test for Facilitation Effects. *American Journal of Public Health*, 103(9):1604–1609.

Drummond, C. (2009). Replicability is not Reproducibility: Nor is it Good Science. In *Proceedings of the Evaluation Methods for Machine Learning Workshop at the 26th ICML*.

DuHigg, C. (2012). Psst, You in Aisle 5. *The New York Times Magazine*, p. MM30. Dumanovsky, T., Huang, C. Y., Nonas, C. A., Matte, T. D., Bassett, M. T., and Silver, L. D. (2011). Changes in energy content of lunchtime purchases from fast food restaurants after introduction of calorie labelling: Cross sectional customer surveys. *BMJ*, 343:d4464.

Dwyer, M. (2013). Coffee drinking tied to lower risk of suicide. *Harvard Gazette*.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press, Cambridge.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, Cambridge.

Eichler, M. (2010). Graphical Gaussian Modelling of Multivariate Time Series with Latent Variables. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Einstein, A., Podolsky, B., and Rosen, N. (1935). Can Quantum-Mechanical Description of Physical Reality Be Considered Complete? *Physical Review*, 47(10):777–780.

Elbel, B., Kersh, R., Brescoll, V. L., and Dixon, L. B. (2009). Calorie Labeling And Food Choices: A First Look At The Effects On Low-Income People In New York City. *Health Affairs*, 28(6):w1110–w1121.

Ellison, B., Lusk, J. L., and Davis, D. (2014). The Effect of Calorie Labels on Caloric Intake and Restaurant Revenue: Evidence from Two Full-Service Restaurants. *Journal of Agricultural and Applied Economics*, 46(2):173–191.

Entner, D. and Hoyer, P. O. (2010). On Causal Discovery from Time Series Data using FCI. In *Proceedings of the 5th European Workshop on Probabilistic Graphical Models*.

Erickson, J. D. (1978). Down syndrome, paternal age, maternal age and birth order. *Annals of Human Genetics*, 41(3):289–298.

Erlwein, O., Kaye, S., McClure, M. O., Weber, J., Wills, G., Collier, D., Wessely, S., and Cleare, A. (2010). Failure to Detect the Novel Retrovirus XMRV in Chronic Fatigue Syndrome. *PloS ONE*, 5(1):e8519.

Faro, D., McGill, A. L., and Hastie, R. (2013). The influence of perceived causation on judgments of time: An integrative review and implications for decisionmaking. *Frontiers in Psychology*, 4:217.

Fewtrell, M. S., Kennedy, K., Singhal, A., Martin, R. M., Ness, A., Hadders-Algra, M., Koletzko, B., and Lucas, A. (2008). How much loss to follow-up is acceptable in long-term randomised trials and prospective studies? *Archives of Disease in Childhood*, 93(6):458–461.

Fischer, D. A. (1992). Causation in Fact in Omission Cases. *Utah Law Review*, pp. 1335–1384.

——— (2006). Insufficient Causes. *Kentucky Law Journal*, 94:277–37.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh.

Fitelson, B. and Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. M. Illari, F. Russo, and J. Williamson (eds.), *Causality in the Sciences*, pp. 600–627. Oxford University Press, Oxford.

Fleming, P. J., Blair, P., Bacon, P., and Berry, J. (eds.) (2000). *Sudden unexpected deaths in infancy: The CESDI SUDI studies 1993–1996*. The Stationery Office, London.

Fowler, J. H. and Dawes, C. T. (2008). Two Genes Predict Voter Turnout. *The Journal of Politics*, 70(3):579–594.

Frank, S. A., Wilson, R., Holloway, R. G., Zimmerman, C., Peterson, D. R., Kieburtz, K., and Kim, S. Y. H. (2008). Ethics of sham surgery: Perspective of patients. *Movement Disorders*, 23(1):63–68.

Freedman, D. and Humphreys, P. (1999). Are There Algorithms That Discover Causal Structure? *Synthese*, 121(1-2):29–54.

Fugelsang, J. A. and Thompson, V. A. (2003). A dual-process model of belief and evidence interactions in causal reasoning. *Memory & Cognition*, 31(5):800–815.

Fumerton, R. and Kress, K. (2001). Causation and the Law: Preemption, Lawful Sufficiency, and Causal Sufficiency. *Law and Contemporary Problems*, 64(4):83–105.

Gabriel, A. and Mercado, C. P. (2011). Data retention after a patient withdraws consent in clinical trials. *Open Access Journal of Clinical Trials*, 3:15–19.

Gemelli, A. and Cappellini, A. (1958). The influence of the subject's attitude in perception. *Acta Psychologica*, 14:12–23.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014.

Glasgow, R. E., Lichtenstein, E., and Marcus, A. C. (2003). Why Don't We See More Translation of Health Promotion Research to Practice? Rethinking the Efficacy-to- Effectiveness Transition. *American Journal of Public Health*, 93(8):1261–1267.

Glennan, S. (1996). Mechanisms and the Nature of Causation. *Erkenntnis*, 44(1): 49–71.

——— (2002). Rethinking Mechanistic Explanation. *Philosophy of Science*, 69(3): S342–S353.

Godfrey-Smith, P. (2010). Causal Pluralism. In H. Beebee, C. R. Hitchcock, and P. Menzies (eds.), *Oxford Handbook of Causation*, pp. 326–337. Oxford University Press, Oxford.

Goldman, D. P., Joyce, G. F., and Zheng, Y. (2007). Prescription Drug Cost Sharing: Associations With Medication and Medical Utilization and Spending and Health. *Journal of the American Medical Association*, 298(1):61–69.

Good, I. J. (1961). A Causal Calculus (I). *British Journal for the Philosophy of Science*, 11(44):305–318.

Gopnik, A., Sobel, D. M., Schulz, L. E., and Glymour, C. (2001). Causal Learning Mechanisms in Very Young Children: Two-, Three-, and Four-Year-Olds Infer Causal Relations From Patterns of Variation and Covariation. *Developmental Psychology*, 37(5):620–629.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., and Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review*, 111(1):3–32.

Granger, C. W. J. (1980). Testing for Causality: A Personal Viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352.

Green, J. (2012). The Science Behind Those Obama Campaign E-Mails. *Bloomberg Businessweek*.

Greville, W. J. and Buehner, M. J. (2010). Temporal Predictability Facilitates Causal Learning. *Journal of Experimental Psychology: General*, 139(4):756–771.

Griffiths, T. L., Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2011). Bayes and Blickets: Effects of Knowledge on Causal Induction in Children and Adults. *Cognitive Science*, 35(8):1407–1455.

Griffiths, T. L. and Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, 51(4):334–384.

Grodstein, F., Stampfer, M. J., Colditz, G. A., Willett, W. C., Manson, J. E., Joffe, M., Rosner, B., Fuchs, C., Hankinson, S. E., Hunter, D. J., Hennekens, C. H., and Speizer, F. E. (1997). Postmenopausal Hormone Therapy and Mortality. *The New England Journal of Medicine*, 336(25):1769–1775.

Groves, R. M., Fowler Jr, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009). *Survey Methodology*, 2nd edition. John Wiley & Sons, Hoboken, NJ 2nd edition.

Grünbaum, A. (1981). The placebo concept. *Behaviour Research and Therapy*, 19(2):157–167.

Grzegorczyk, M. and Husmeier, D. (2009). Non-stationary continuous dynamic Bayesian networks. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*.

Guyatt, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., and Schünemann, H. J. (2008). GRADE: An emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*, 336(7650):924–926.

Gweon, H. and Schulz, L. (2011). 16-Month-Olds Rationally Infer Causes of Failed Actions. *Science*, 332(6037):1524.

Hajjar, E. R., Cafiero, A. C., and Hanlon, J. T. (2007). Polypharmacy in elderly patients. *The American Journal of Geriatric Pharmacotherapy*, 5(4):345–351.

Halpern, J. Y. and Hitchcock, C. R. (2010). Actual Causation and the Art of Modeling. In R. Dechter, H. Geffner, and J. Y. Halpern (eds.), *Heuristics, Probability and Causality: A Tribute to Judea Pearl*, pp. 383–406. College Publications, London.

Hanushek, E. A. (2011). The economic value of higher teacher quality. *Economics of Education Review*, 30(3):466–479.

Hart, H. L. A. and Honoré, T. (1985). *Causation in the Law*. Oxford University Press, Oxford.

Haskins, R. and Sawhill, I. V. (2009). *Creating an Opportunity Society*. Brookings Institution Press, Washington, DC.

Hastie, R. and Pennington, N. (1996). The O.J. Simpson Stories: Behavioral Scientists' Reflections on The People of the State of California v. Orenthal James Simpson. *University of Colorado Law Review*, 67:957–976.

Haushofer, J. and Shapiro, J. (2013). Household response to income changes: Evidence from an unconditional cash transfer program in Kenya. Technical report.

Hausman, D. M. (2005). Causal Relata: Tokens, Types, or Variables? *Erkenntnis*, 63(1):33–54.

Heeger, D. J. and Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews Neuroscience*, 3(2):142–151.

Heider, F. and Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, 57(2):243–259.

Henning, A. S. (1948). Dewey defeats Truman. *Chicago Tribune* p. 1.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2-3):61–83.

Heres, S., Davis, J., Maino, K., Jetzinger, E., Kissling, W., and Leucht, S. (2006). Why Olanzapine Beats Risperidone, Risperidone Beats Quetiapine, and Quetia- pine Beats Olanzapine: An Exploratory Analysis of Head-to-Head Comparison Studies of Second-Generation Antipsychotics. *American Journal of Psychiatry*, 163(2):185–194.

Hernan, M. A., Clayton, D., and Keiding, N. (2011). The Simpson's paradox unraveled. *International Journal of Epidemiology*, 40(3):780–785.

Herndon, T., Ash, M., and Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2):257–279.

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300.

Hitchcock, C. and Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 106(11): 587–612.

Hitchcock, C. R. (1995). The Mishap at Reichenbach Fall: Singular vs. General Causation. *Philosophical Studies*, 78(3):257–291.

Ho, Y. C. and Pepyne, D. L. (2002). Simple Explanation of the No-Free-Lunch Theorem and Its Implications. *Journal of Optimization Theory and Applications*, 115(3):549–570.

Höfler, T., Przyrembel, H., and Verleger, S. (2004). New evidence for the Theory of the Stork. *Paediatric and Perinatal Epidemiology*, 18(1):88–92.

Höfler, M. (2005). The Bradford Hill considerations on causality: A counterfactual perspective. *Emerging Themes in Epidemiology*, 2:11.

Holgate, S. T., Komaroff, A. L., Mangan, D., and Wessely, S. (2011). Chronic fatigue syndrome: Understanding a complex illness. *Nature Reviews Neuroscience*, 12(9): 539–44.

Holson, L. M. (2009). Putting a Bolder Face on Google. *The New York Times* p. B1.

Holyoak, K. J. and Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology*, 62:135–163.

Howick, J. (2011). *Placebo Controls: Problematic and Misleading Baseline Measures of Effectiveness*, pp. 80–95. Wiley-Blackwell, Chichester, West Sussex, UK.

Howick, J., Chalmers, I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., Moschetti, I., Phillips, B., and Thornton, H. (2011). Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document).

Howick, J., Glasziou, P., and Aronson, J. K. (2009). The evolution of evidence hierarchies: What can Bradford Hill's 'guidelines for causation' contribute? *The Journal of the Royal Society of Medicine*, 102(5):186–194.

Hripcsak, G., Elhadad, N., Chen, Y. H., Zhou, L., and Morrison, F. P. (2009). Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts. *Journal of the American Medical Informatics Association*, 16(2):220–227.

Hué, S., Gray, E. R., Gall, A., Katzourakis, A., Tan, C. P., Houldcroft, C. J., McLaren, S., Pillay, D., Futreal, A., and Garson, J. A. (2010). Disease-associated XMRV sequences are consistent with laboratory contamination. *Retrovirology*, 7(1):111.

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., and Vittinghoff, E. (1998). Randomized Trial of Estrogen Plus Progestin for Secondary Prevention of Coronary Heart Disease in Postmenopausal Women. *JAMA*, 280(7):605–613.

Hume, D. (1739). *A Treatise of Human Nature*. London. Reprint, Prometheus Books, 1992. Citations refer to the Prometheus edition.

——— (1748). *An Enquiry Concerning Human Understanding*. London. Reprint, Dover Publications, 2004.

Illari, P. and Russo, F. (2014). *Causality: Philosophical Theory Meets Scientific Practice*. Oxford University Press, Oxford.

Issenberg, S. (2012). *The Victory Lab: The Secret Science of Winning Campaigns*. Crown, New York.

Jepsen, C. and Rivkin, S. (2009). Class Reduction and Student Achievement: The Potential Tradeoff between Teacher Quality and Class Size. *Journal of Human Resources*, 44(1):223–250.

Johnson, S. R. (2008). The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *Journal of Chemical Information and Modeling*, 48(1):25–26.

Joynson, R. B. (1971). Michotte's experimental methods. *British Journal of Psychology*, 62(3):293–302.

Kahneman, D., Lovallo, D., and Sibony, O. (2011). Before You Make That Big Decision... *Harvard Business Review*, 89(6):50–60.

Kant, I. (1902). *Prolegomena to Any Future Metaphysics*. Open Court Publishing, Chicago. Translated by Paul Carus.

——— (1998). *Critique of Pure Reason*. Cambridge University Press, Cambridge. Translated by Paul Guyer and Allen W. Wood.

Kaptchuk, T. J., Friedlander, E., Kelley, J. M., Sanchez, M. N., Kokkotou, E., Singer, J. P., Kowalczykowski, M., Miller, F. G., Kirsch, I., and Lembo, A. J. (2010). Placebos without Deception: A Randomized Controlled Trial in Irritable Bowel Syndrome. *PloS ONE*, 5(12):e15591.

Kearney, M. S. and Levine, P. B. (2014). Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing. Technical Report 19795, National Bureau of Economic Research.

Keeter, S., Dimock, M., and Christian, L. (2008). Calling Cell Phones in '08 Pre- Election Polls. *The Pew Research Center for the People and the Press*.

Kiszko, K. M., Martinez, O. D., Abrams, C., and Elbel, B. (2014). The Influence of Calorie Labeling on Food Orders and Consumption: A Review of the Literature. *Journal of Community Health*, 39(6):1248–1269.

Klein, R. A., Ratliff, K. A., Vianello, M., et al. (2014). Investigating Variation in Replicability. *Social Psychology*, 45(3):142–152.

Kleinberg, S. (2012). *Causality, Probability, and Time*. Cambridge University Press, New York.

Kleinberg, S. and Elhadad, N. (2013). Lessons Learned in Replicating Data-Driven Experiments in Multiple Medical Systems and Patient Populations. In *AMIA Annual Symposium*.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63(279):190–194.

Knobe, J. and Fraser, B. (2008). Causal Judgment and Moral Judgment: Two Experiments. In W. Sinnott-Armstrong (ed.), *Moral Psychology*, volume 2, pp.441–448. The MIT Press, Cambridge, MA.

Knobe, J. and Mendlow, G. S. (2004). The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology. *Journal of Theoretical and Philosophical Psychology*, 24(2):252–258.

Knobe, J. and Nichols, S. (2008). *Experimental Philosophy*. Oxford University Press, Oxford.

Koch, R. (1932). Die Aetiologie der Tuberkulose. *Journal of Molecular Medicine*, 11(12):490–492.

Koppett, L. (1978). Carrying Statistics to Extremes. *Sporting News*.

Korja, M., Silventoinen, K., Laatikainen, T., Jousilahti, P., Salomaa, V., Hernesniemi, J., and Kaprio, J. (2013). Risk Factors and Their Combined Effects on the Incidence Rate of Subarachnoid Hemorrhage – A Population-Based Cohort Study. *PLoS ONE*, 8(9):e73760.

Kravitz, R. L. and Duan, N. (eds.) (2014). *Design and Implementation of N-of-1 Trials: A User's Guide*. Agency for Healthcare Research and Quality, Rockville, MD.

Krieger, J. and Saelens, B. E. (2013). Impact of Menu Labeling on Consumer Behavior: A 2008–2012 Update. *Robert Wood Johnson Foundation*.

Krueger, A. B. (2003). Economic Considerations and Class Size. *The Economic Journal*, 113(485):F34–F63.

Krueger, A. B. and Whitmore, D. M. (2001). The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR. *The Economic Journal*, 111(468):1–28.

van Kuppeveld, F. J., de Jong, A. S., Lanke, K. H., Verhaegh, G. W., Melchers, W. J., Swanink, C. M., Bleijenberg, G., Netea, M. G., Galama, J. M., and van Der Meer, J. W. (2010). Prevalence of xenotropic murine leukaemia virus-related virus in patients with chronic fatigue syndrome in the Netherlands: Retrospective analysis of samples from an established cohort. *BMJ*, 340:c1018.

Kushnir, T. and Gopnik, A. (2005). Young Children Infer Causal Strength from Probabilities and Interventions. *Psychological Science*, 16(9):678–683.

Kwiatkowski, D., Phillips, P. C., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178.

Lagnado, D. A. and Channon, S. (2008). Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition*, 108(3):754–770.

Lagnado, D. A. and Harvey, N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review*, 15(6):1166–1173.

Lagnado, D. A. and Sloman, S. (2004). The Advantage of Timely Intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30(4):856–876.

Lagnado, D. A. and Sloman, S. A. (2006). Time as a Guide to Cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3):451–460.

Lagnado, D. A. and Speekenbrink, M. (2010). The Influence of Delays in Real-Time Causal Learning. *The Open Psychology Journal*, 3(2):184–195.

Lagnado, D. A., Waldmann, M. R., Hagmayer, Y., and Sloman, S. A. (2007). Beyond Covariation. In A. Gopnik and L. Schulz (eds.), *Causal learning: Psychology, Philosophy, and Computation*, pp. 154–172. Oxford University Press, Oxford.

Lange, M. (2013). What Makes a Scientific Explanation Distinctively Mathematical? *The British Journal for the Philosophy of Science*, 64(3):485–511.

Lazer, D. M., Kennedy, R., King, G., and Vespignani, A. (2014). The Parable of Google Flu: Traps in Big Data Analysis. *Science*, 343(6176):1203–1205.

Leibovici, L. (2001). Effects of remote, retroactive intercessory prayer on outcomes in patients with bloodstream infection: Randomised controlled trial. *BMJ*, 323(7327): 1450–1451.

Leslie, A. M. (1982). The perception of causality in infants. *Perception*, 11(2):173–186.

Leslie, A. M. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, 70(17):556–567. Reprinted in Lewis 1986a.

——— (1976). The paradoxes of time travel. *American Philosophical Quarterly*, 13(2):145–152.

——— (1986a). *Philosophical Papers*, volume 2. Oxford University Press, Oxford.

——— (1986b). Postscripts to "Causation". In *Philosophical Papers*, volume 2, pp. 172–213. Oxford University Press, Oxford.

——— (2000). Causation as Influence. *The Journal of Philosophy*, 97(4):182–197.

Lin, P. and Gill, J. R. (2009). Delayed Homicides and the Proximate Cause. *American Journal of Forensic Medicine & Pathology*, 30(4):354–357.

Lind, J. (1757). *A Treatise on the Scurvy: In Three Parts, Containing an Inquiry Into the Nature, Causes, and Cure, of that Disease*. A. Millar, London.

Linthwaite, S. and Fuller, G. N. (2013). Milk, chocolate and Nobel prizes. *Practical Neurology*, 13(1):63–63.

Lo, S.-C., Pripuzova, N., Li, B., Komaroff, A. L., Hung, G.-C., Wang, R., and Alter, H. J. (2010). Detection of MLV-related virus gene sequences in blood of patients with chronic fatigue syndrome and healthy blood donors. *Proceedings of the National Academy of Sciences*, 107(36):15874–15879.

——— (2012). Retraction for Lo et al., Detection of MLV-related virus gene sequences in blood of patients with chronic fatigue syndrome and healthy blood donors. *Proceedings of the National Academy of Sciences*, 109(1):346–346.

Lombardi, V. C., Ruscetti, F. W., Gupta, J. D., Pfost, M. A., Hagen, K. S., Peterson, D. L., Ruscetti, S. K., Bagni, R. K., Petrow-Sadowski, C., Gold, B., Dean, M., Silverman, R. H., and Mikovits, J. A. (2009). Detection of an Infectious Retrovirus, XMRV, in Blood Cells of Patients with Chronic Fatigue Syndrome. *Science*, 326(5952):585–589.

Lopes, L. (1993). Two conceptions of the juror. In R. Hastie (ed.), *Inside the Juror: The Psychology of Juror Decision Making*, pp. 255–262. Cambridge University Press, Cambridge.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, 67(1):1–25.

Mackie, J. L. (1974). *The Cement of the Universe*. Clarendon Press, Oxford.

Macklin, R. (1999). The Ethical Problems with Sham Surgery in Clinical Research. *The New England Journal of Medicine*, 341(13):992–996.

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). A Theory of Blame. *Psychological Inquiry: An International Journal for the Advancement of Psychological Theory*, 25(2):147–186.

Mandel, D. R. (2003). Judgment Dissociation Theory: An Analysis of Differences in Causal, Counterfactual, and Covariational Reasoning. *Journal of Experimental Psychology: General*, 132(3):419–434.

March, L., Irwig, L., Schwarz, J., Simpson, J., Chock, C., and Brooks, P. (1994). n of 1 trials comparing a non-steroidal anti-inflammatory drug with paracetamol in osteoarthritis. *BMJ*, 309(6961):1041–1045.

Matossian, M. A. K. (1989). *Poisons of the Past: Molds, Epidemics, and History*. Yale University Press, New Haven, CT.

Matthews, R. (2000). Storks Deliver Babies (p=0.008). *Teaching Statistics*, 22(2): 36–38.

Maurage, P., Heeren, A., and Pesenti, M. (2013). Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies. *The Journal of Nutrition*, 143(6):931–933.

McLean, K. A., Byanaku, A., Kubikonse, A., Tshowe, V., Katensi, S., and Lehman, A. G. (2014). Fishing with bed nets on Lake Tanganyika: A randomized survey. *Malaria Journal*, 13:395.

McLean, R. D. and Pontiff, J. (2015). Does Academic Research Destroy Stock Return Predictability? *Journal of Finance*, forthcoming.

Meadow, R. (2002). A case of murder and the BMJ. *BMJ*, 324(7328):41–43.

Meek, C. (1995). Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*.

Meeks, R. R. (2004). Unintentionally Biasing the Data: Reply to Knobe. *Journal of Theoretical and Philosophical Psychology*, 24(2):220–223.

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *The New England Journal of Medicine*, 367(16):1562–1564.

Michotte, A. (1946). *La Perception de la Causalité*. Editions de l'Institut Supérieur de Philosophie, Louvain. English translation by T. Miles & E. Miles. *The Perception of Causality*, Basic Books, 1963. Citations refer to the translated edition.

Mill, J. S. (1843). *A System of Logic*. Parker, London. Reprint, Lincoln-Rembrandt Pub., 1986.

Miller, J. G. (1984). Culture and the Development of Everyday Social Explanation. *Journal of Personality and Social Psychology*, 46(5):961–978.

Mitofsky, W. J. (1998). Review: Was 1996 a Worse Year for Polls Than 1948? *The Public Opinion Quarterly*, 62(2):230–249.

Mitsumoto, H., Brooks, B. R., and Silani, V. (2014). Clinical trials in amyotrophic lateral sclerosis: Why so many negative trials and how can trials be improved? *The Lancet Neurology*, 13(11):1127–1138.

Moher, D., Schulz, K. F., and Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, 357(9263):1191–1194.

Morris, M. W. and Peng, K. (1994). Culture and Cause: American and Chinese Attributions for Social and Physical Events. *Journal of Personality and Social Psychology*, 67(6):949–971.

Mosca, L., Manson, J. E., Sutherland, S. E., Langer, R. D., Manolio, T., and Barrett-Connor, E. (1997). Cardiovascular disease in women: A statement for healthcare professionals from the American Heart Association. Writing Group. *Circulation*, 96(7):2468–2482.

Mostofsky, E., Rice, M. S., Levitan, E. B., and Mittleman, M. A. (2012). Habitual Coffee Consumption and Risk of Heart Failure: A Dose-Response Meta-Analysis. *Circulation: Heart Failure*, 5(4):401–405.

Mott, N. L. (2003). The Current Debate on Juror Questions: To Ask or Not to Ask, That Is the Question. *Chicago-Kent Law Review*, 78:1099.

Muntner, P., Safford, M. M., Cushman, M., and Howard, G. (2014). Comment on the Reports of Over-estimation of ASCVD Risk Using the 2013 AHA/ACC Risk Equation. *Circulation*, 129(2):266–267.

Murphy, K. (2002). *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, University of California, Berkley.

Nadelhoffer, T. (2004). On Praise, Side Effects, and Folk Ascriptions of Intentionality. *Journal of Theoretical and Philosophical Psychology*, 24(2):196–213.

Narayanan, A. and Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the IEEE Symposium on Security and Privacy*.

Newburger, J. W., Takahashi, M., Gerber, M. A., Gewitz, M. H., Tani, L. Y., Burns, J. C., Shulman, S. T., Bolger, A. F., Ferrieri, P., Baltimore, R. S., Wilson, W. R., Baddour, L. M., Levison, M. E., Pallasch, T. J., Falace, D. A., and Taubert, K. A. (2004). Diagnosis, Treatment, and Long-Term Management of Kawasaki Disease. *Circulation*, 110(17):2747–2771.

Nieman, D. C. (1994). Exercise, Infection, and Immunity. *International Journal of Sports Medicine*, 15(S 3):S131–S141.

Norenzayan, A. and Schwarz, N. (1999). Telling what they want to know: Participants tailor causal attributions to researchers' interests. *European Journal of Social Psychology*, 29(8):1011–1020.

Normore, A. H. and Ilon, L. (2006). Cost-Effective School Inputs: Is Class Size Reduction the Best Educational Expenditure for Florida? *Educational Policy*, 20(2):429–454.

Noseworthy, J. H., Ebers, G. C., Vandervoort, M. K., Farquhar, R. E., Yetisir, E., and Roberts, R. (1994). The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*, 44(1):16–20.

Novick, L. R. and Cheng, P. W. (2004). Assessing Interactive Causal Influence. *Psychological Review*, 111(2):455–485.

Oakes, B., Tai, A. K., Cingöz, O., Henefield, M. H., Levine, S., Coffin, J. M., and Huber, B. T. (2010). Contamination of human DNA samples with mouse DNA can lead to false detection of XMRV-like sequences. *Retrovirology*, 7:109.

Oakes, L. M. (1994). Development of Infants' Use of Continuity Cues in Their Perception of Causality. *Developmental Psychology*, 30(6):869–879.

O'Malley, K. J., Cook, K. F., Price, M. D., Wildes, K. R., Hurdle, J. F., and Ashton, C. M. (2005). Measuring Diagnoses: ICD Code Accuracy. *Health Services Research*, 40(5p2):1620–39.

Ou, Z. Y., Pereira, S. F., Kimble, H. J., and Peng, K. C. (1992). Realization of the Einstein-Podolsky-Rosen paradox for continuous variables. *Physics Review Letters*, 68(25):3663–3666.

Paprotka, T., Delviks-Frankenberry, K. A., Cingöz, O., Martinez, A., Kung, H.-J., Tepper, C. G., Hu, W.-S., Fivash, M. J., Coffin, J. M., and Pathak, V. K. (2011). Recombinant origin of the retrovirus XMRV. *Science*, 333(6038):97–101.

Patberg, W. R. and Rasker, J. J. (2004). Weather effects in rheumatoid arthritis: From controversy to consensus. A review. *The Journal of Rheumatology*, 31(7):1327–1334.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge.

——— (2014). Understanding Simpson's Paradox. *The American Statistician*, 68(1):8–13.

Pearson, K., Lee, A., and Bramley-Moore, L. (1899). Mathematical Contributions to the Theory of Evolution. VI. Genetic (Reproductive) Selection: Inheritance of Fertility in Man, and of Fecundity in Thoroughbred Racehorses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 192:257–330.

Peng, K. and Knowles, E. D. (2003). Culture, Education, and the Attribution of Physical Causality. *Personality and Social Psychology Bulletin*, 29(10):1272–1284.

Pennington, N. and Hastie, R. (1986). Evidence Evaluation in Complex Decision Making. *Journal of*

Personality and Social Psychology, 51(2):242–258.

——— (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(3):521–533.

——— (1992). Explaining the Evidence: Tests of the Story Model for Juror Decision Making. *Journal of Personality and Social Psychology*, 62(2):189–206.

Perales, J. C., Shanks, D. R., and Lagnado, D. (2010). Causal Representation and Behavior: The Integration of Mechanism and Covariation. *Open Psychology Journal*, 3(1):174–183.

Perotte, A. and Hripcsak, G. (2013). Temporal Properties of Diagnosis Code Time Series in Aggregate. *IEEE Journal of Biomedical and Health Informatics*, 17(2): 477–483.

Perwien, A. R., Johnson, S. B., Dymtrow, D., and Silverstein, J. (2000). Blood Glucose Monitoring Skills in Children with Type I Diabetes. *Clinical Pediatrics*, 39(6):351–357.

Phillips, C. V. and Goodman, K. J. (2004). The missed lessons of Sir Austin Bradford Hill. *Epidemiologic Perspectives & Innovations*, 1(1):3.

Pivovarov, R. and Elhadad, N. (2012). A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of Biomedical Informatics*, 45(3):471–481.

Power, D. J. (2002). Ask Dan! What is the “true story” about data mining, beer and diapers? *DSS News*, 3(23).

Price, D. D., Finniss, D. G., and Benedetti, F. (2008). A Comprehensive Review of the Placebo Effect: Recent Advances and Current Thought. *Annual Review of Psychology*, 59:565–590.

Price, H. (1997). *Time’s Arrow and Archimedes’ Point: New Directions for the Physics of Time*. Oxford University Press, Oxford.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9):712–713.

Pritchard, C. (2012). Does chocolate make you clever? BBC News.

Pronin, E., Wegner, D. M., McCarthy, K., and Rodriguez, S. (2006). Everyday Magical Powers: The Role of Apparent Mental Causation in the Overestimation of Personal Influence. *Journal of Personality and Social Psychology*, 91(2):218–231.

Psillos, S. (2010). Causal Pluralism. In R. Vanderbeeken and B. D’Hooghe (eds.), *World-views, Science and Us: Studies of Analytical Metaphysics*, pp. 131–151. World Scientific Publishers, Singapore.

R v. Jordan (1956). 40 Cr App R. 152.

Radelet, M. L. and Pierce, G. L. (1991). Choosing Those Who Will Die: Race and the Death Penalty in Florida. *Florida Law Review*, 43(1):1–34.

Redelmeier, D. A. and Tversky, A. (1996). On the belief that arthritis pain is related to the weather. *Proceedings of the National Academy of Sciences*, 93(7):2895–2896.

Reichenbach, H. (1956). *The Direction of Time*. University of California Press, Berkeley. Reprint, Dover Publications, 2000.

Reiss, J. (2007). Time Series, Nonsense Correlations and the Principle of the Common Cause. In F. Russo and J. Williamson (eds.), *Causality and Probability in the Sciences*, pp. 179–196. College Publications, London.

——— (2014). What’s Wrong With Our Theories of Evidence? *Theoria*, 29(2): 283–306.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black and W. F. Prokasy (eds.), *Classical Conditioning II: Current Theory and Research*, pp. 64–99. Appleton-Century-Crofts, New York.

Rhonheimer, J. (writer) and Fryman, P. (director). (2007). Lucky penny [Television series episode]. In Bays, C. and Thomas, C. (producers), *How I met your mother*. CBS, Los Angeles.

Ridker, P. M. and Cook, N. R. (2013). Statins: New American guidelines for prevention of cardiovascular disease. *The Lancet*, 382(9907):1762–1765.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity Analysis for Selection bias and unmeasured Confounding in missing Data and Causal inference models. In M. E. Halloran and D. Berry (eds.), *Statistical Models in Epidemiology: The Environment and Clinical Trials*, pp. 1–94. Springer-Verlag, New York.

Robinson, J. W. and Hartemink, A. J. (2010). Learning Non-Stationary Dynamic Bayesian Networks. *Journal of Machine Learning Research*, 11(Dec):3647–3680.

Robinson, M. J., Erlwein, O. W., Kaye, S., Weber, J., Cingoz, O., Patel, A., Walker, M. M., Kim, W.-J. J., Uiprasertkul, M., Coffin, J. M., and McClure, M. O. (2010). Mouse DNA contamination in human tissue tested for XMRV. *Retrovirology*, 7:108.

de Rooij, N. K., Linn, F. H. H., van der Plas, J. A., Algra, A., and Rinkel, G. J. E. (2007). Incidence of subarachnoid haemorrhage: A systematic review with emphasis on region, age, gender and time trends. *Journal of Neurology, Neurosurgery & Psychiatry*, 78(12):1365–72.

Roser, M. E., Fugelsang, J. A., Dunbar, K. N., Corballis, P. M., and Gazzaniga, M. S. (2005). Dissociating Processes Supporting Causal Perception and Causal Inference in the Brain. *Neuropsychology*, 19(5):591–602.

Rothman, K. J. (1976). Causes. *American Journal of Epidemiology*, 104(6):587–592. Reprinted in 141(2), 1995.

——— (1990). No Adjustments Are Needed for Multiple Comparisons. *Epidemiology*, 1(1):43–46.

Rothman, K. J. and Greenland, S. (2005). Causation and Causal Inference in Epidemiology. *American Journal of Public Health*, 95(S1):S144–S150.

Rothwell, P. M. (2005). External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*, 365(9453):82–93.

Russell, B. (1912). On the Notion of Cause. *Proceedings of the Aristotelian Society*, 13(1912-1913):1–26.

Russo, F. (2006). The Rationale of Variation in Methodological and Evidential Pluralism. *Philosophica*, 77(1):97–124.

Russo, F. and Williamson, J. (2007). Interpreting Causality in the Health Sciences. *International Studies in the Philosophy of Science*, 21(2):157–170.

Salganik, M. J., Dodds, P. S., and Watts, D. J. (2006). Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science*, 311(5762):854–856.

Sandvei, M., Mathiesen, E., Vatten, L., Müller, T., Lindekleiv, H., Ingebrigtsen, T., Njøstad, I., Wilsgaard, T., Løhen, M.-L., Vik, A., et al. (2011). Incidence and mortality of aneurysmal subarachnoid hemorrhage in two Norwegian cohorts, 1984–2007. *Neurology*, 77(20):1833–1839.

Sato, E., Furuta, R. A., and Miyazawa, T. (2010). An Endogenous Murine Leukemia Viral Genome Contaminant in a Commercial RT-PCR Kit is Amplified Using Standard Primers for XMRV. *Retrovirology*, 7(1):110.

Saunders System Birmingham Co. v. Adams (1928). 217 Ala. 621, 117 So. 72.

Scheines, R. (1997). An Introduction to Causal Inference. In V. R. McKim and S. P. Turner (eds.), *Causality in Crisis*, pp. 185–199. University of Notre Dame Press, Notre Dame, IN.

Schlottmann, A. (1999). Seeing It Happen and Knowing How It Works: How Children Understand the Relation Between Perceptual Causality and Underlying Mechanism. *Developmental Psychology*,

35(5):303–317.

Schlottmann, A., Allen, D., Linderoth, C., and Hesketh, S. (2002). Perceptual Causality in Children. *Child Development*, 73(6):1656–1677.

Schlottmann, A., Ray, E. D., and Surian, L. (2012). Emerging perception of causality in action-and-reaction sequences from 4 to 6 months of age: Is it domainspecific? *Journal of Experimental Child Psychology*, 112(2):208–230.

Schlottmann, A. and Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *The Quarterly Journal of Experimental Psychology*, 44(2):321–342.

Schoenfeld, J. D. and Ioannidis, J. P. (2013). Is everything we eat associated with cancer? A systematic cookbook review. *The American Journal of Clinical Nutrition*, 97(1):127– 134.

Schulz, K. F. and Grimes, D. A. (2002). Blinding in randomised trials: Hiding who got what. *The Lancet*, 359(9307):696–700.

Schulz, L. E., Gopnik, A., and Glymour, C. (2007). Preschool children learn about causal structure from conditional interventions. *Developmental Science*, 10(3):322–332.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Scriven, M. (1966). Causes, connections and conditions in history. In W. H. Dray (ed.), *Philosophical Analysis and History*, pp. 238–264. Harper & Row, New York.

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., et al. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 110(9):3507–3512.

Shalizi, C. R. and Thomas, A. C. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods Research*, 40(2):211–239.

Shanks, D. R. (1985). Forward and backward blocking in human contingency judgement. *The Quarterly Journal of Experimental Psychology*, 37(1):1–21.

——— (1995). *The Psychology of Associative Learning*. Cambridge University Press, Cambridge.

Shanks, D. R., Pearson, S. M., and Dickinson, A. (1989). Temporal Contiguity and the Judgement of Causality by Human Subjects. *The Quarterly Journal of Experimental Psychology*, 41 B(2):139–159.

Sidhu, D. (2015). Moneyball Sentencing. *Boston College Law Review*, 56(2):671–731.

Silverman, R. H., Das Gupta, J., Lombardi, V. C., Ruscetti, F. W., Pfost, M. A., Hagen, K. S., Peterson, D. L., Ruscetti, S. K., Bagni, R. K., Petrow-Sadowski, C., Gold, B., Dean, M., and Mikovits, J. (2011). Partial retraction. *Science*, 334(6053):176.

Simpson, E. H. (1951). The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 13(2):238–241.

Skyrms, B. (1984). EPR: Lessons for Metaphysics. *Midwest Studies in Philosophy*, 9(1):245–255.

Slobogin, C. (2012). Risk Assessment. In J. Petersilia and K. R. Reitz (eds.), *Oxford Handbook of Sentencing and Corrections*, pp. 196–214. Oxford University Press, New York.

Sloman, S. A. and Lagnado, D. (2015). Causality in Thought. *Annual Review of Psychology*, 66:223–247.

Smith, G. C. S. and Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: Systematic review of randomised controlled trials. *BMJ*, 327(7429):1459–1461.

Snow, J. (1854). The Cholera Near Golden Square, and at Deptford. *Medical Times and Gazette*, 9:321–322.

——— (1855). *On the Mode of Communication of Cholera*. John Churchill, London.

Sobel, D. M. and Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology*, 42(6):1103–1115.

Sobel, D. M. and Kushnir, T. (2006). The importance of decision making in causal learning from interventions. *Memory & Cognition*, 34(2):411–419.

Sobel, D. M., Tenenbaum, J. B., and Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, 28(3):303–333.

Sober, E. (1987). Parsimony, Likelihood, and the Principle of the Common Cause. *Philosophy of Science*, 54(3):465–469.

——— (2001). Venetian Sea Levels, British Bread Prices, and the Principle of the Common Cause. *British Journal for the Philosophy of Science*, 52(2):331–346.

Sober, E. and Papineau, D. (1986). Causal Factors, Causal Inference, Causal Explanation. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 60:97–136.

Sonnenberg, L., Gelsomin, E., Levy, D. E., Riis, J., Barraclough, S., and Thorndike, A. N. (2013). A traffic light food labeling intervention increases consumer awareness of health and healthy choices at the point-of-purchase. *Preventive Medicine*, 57(4):253–257.

Spanos, N. P. and Gottlieb, J. (1976). Ergotism and the Salem Village Witch Trials. *Science*, 194(4272):1390–1394.

Spellman, B. A. (1996). Acting as Intuitive Scientists: Contingency Judgments Are Made while Controlling for Alternative Potential Causes. *Psychological Science*, 7(6):337–342.

Spellman, B. A. and Kincannon, A. (2001). The Relation between Counterfactual ("But for") and Causal Reasoning: Experimental Findings and Implications for Jurors' Decisions. *Law and Contemporary Problems*, 64(4):241–264.

Spencer, S. J., Steele, C. M., and Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, 35(1):4–28.

Spirtes, P. (2005). Graphical models, causal inference, and econometric models. *Journal of Economic Methodology*, 12(1):3–34.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*, 2nd edition. The MIT Press, Cambridge, MA. First published 1993.

Spirtes, P., Meek, C., and Richardson, T. (1995). Causal Inference in the Presence of Latent Variables and Selection Bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*.

Starr, S. B. (2014). Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review*, 66:803.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., and Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3):453–489.

Stone, N. J., Robinson, J., Lichtenstein, A. H., Merz, C. N. B., Blum, C. B., Eckel, R. H., Goldberg, A. C., Gordon, D., Levy, D., Lloyd-Jones, D. M., McBride, P., Schwartz, J. S., Shero, S. T., Smith, S. C., Watson, K., and Wilson, P. W. (2013). 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Journal of the American College of Cardiology*, 63(25):2889–2934.

Stoppard, T. (director). (1990). *Rosencrantz & Guildenstern Are Dead* [Motion picture]. Cinecom Pictures, New York.

Subbotsky, E. (2004). Magical thinking in judgments of causation: Can anomalous phenomena affect ontological causal beliefs in children and adults? *British Journal of Developmental Psychology*, 22(1):123–152.

Sudman, S. and Blair, E. (1999). Sampling in the Twenty-First Century. *Journal of the Academy of Marketing Science*, 27(2):269–277.

Sullivan, W. (1982). New Study Backs Thesis on Witches. *The New York Times* p. 30.

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland, Amsterdam.

Susser, M. (1991). What is a Cause and How Do We Know One? A Grammar for Pragmatic Epidemiology. *American Journal of Epidemiology*, 133(7):635–648.

Swartz, J. J., Braxton, D., and Viera, A. J. (2011). Calorie menu labeling on quickservice restaurant menus: An updated systematic review of the literature. *International Journal of Behavioral Nutrition and Physical Activity*, 8(1):135.

Takao, K. and Miyakawa, T. (2014). Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences*, 112(4):1167–1172.

Tatonetti, N. P., Denny, J. C., Murphy, S. N., Fernald, G. H., Krishnan, G., Castro, V., Yue, P., Tsau, P. S., Kohane, I., Roden, D. M., and Altman, R. B. (2011).

Detecting Drug Interactions From Adverse-Event Reports: Interaction Between Paroxetine and Pravastatin Increases Blood Glucose Levels. *Clinical Pharmacology & Therapeutics*, 90(1):133–142.

Thompson, W. C. and Schumann, E. L. (1987). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11(3):167–187.

Thurman, W. N. and Fisher, M. E. (1988). Chickens, Eggs, and Causality, or Which Came First? *American Journal of Agricultural Economics*, 70(2):237–238.

Tulppo, M. P., Hautala, A. J., Mäkikallio, T. H., Laukkanen, R. T., Nissilä, S., Hughson, R. L., and Huikuri, H. V. (2003). Effects of aerobic training on heart rate dynamics in sedentary subjects. *Journal of Applied Physiology*, 95(1):364–372.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., and Rosenthal, R. (2008). Selective Publication of Antidepressant Trials and Its Influence on Apparent Efficacy. *The New England Journal of Medicine*, 358(3):252–260.

Tversky, A. and Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Uttich, K. and Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, 116(1):87–100.

Vandenbroucke, J. P. (2004). When are observational studies as credible as randomised trials? *The Lancet*, 363(9422):1728–1731.

Vickers, A. (2010). *What is a P-value anyway?: 34 stories to help you actually understand statistics*. Addison-Wesley, Boston.

Vlahos, J. (2012). The Case of the Sleeping Slayer. *Scientific American*, 307(3):48–53.

Waldmann, M. R. and Hagmayer, Y. (2005). Seeing Versus Doing: Two Modes of Accessing Causal Knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(2):216–227.

Ward, A. C. (2009). The role of causal criteria in causal inferences: Bradford Hill's "aspects of association." *Epidemiologic Perspectives & Innovations*, 6(1):2.

Watts, D. J. (2011). *Everything Is Obvious: How Common Sense Fails Us*. Crown Business, New York.

Waxman, O. B. (2012). Secret to Winning a Nobel Prize? Eat More Chocolate. *TIME.com*.

Weiss, N. S. (2002). Can the "Specificity" of an Association be Rehabilitated as a Basis for Supporting a Causal Hypothesis? *Epidemiology*, 13(1):6–8.

White, P. (2013). Apportionment of responsibility in medical negligence. *North East Law Review*, 1:147–151.

Wicks, P., Vaughan, T. E., Massagli, M. P., and Heywood, J. (2011). Accelerated clinical discovery using self-reported patient data collected online and a patientmatching algorithm. *Nature Biotechnology*, 29(5):411–414.

Wiener, N. (1956). The theory of prediction. In E. Beckenbach (ed.), *Modern Mathematics for the Engineer*, pp. 165–190. McGraw-Hill, New York.

Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82.

Woodward, J. (2005). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

Woolf, A. (2000). Witchcraft or Mycotoxin? The Salem Witch Trials. *Clinical Toxicology*, 38(4): 457–460.

Wright, R. W. (1985). Causation in Tort Law. *California Law Review*, 73(6):1735–1828.

——— (1987). Causation, Responsibility, Risk, Probability, Naked Statistics, and Proof: Pruning the Bramble Bush by Clarifying the Concepts. *Iowa Law Review*, 73:1001–1077.

——— (2007). Acts and Omissions as Positive and Negative Causes. In J. W. Neyers, E. Chamberlain, and S. G. A. Pitel (eds.), *Emerging Issues in Tort Law*, pp. 287–307. Hart Publishing, Oxford.

Writing Group for the Women's Health Initiative Investigators (2002). Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative Randomized Controlled Trial. *JAMA*, 288(3):321–333.

Young, S. S. and Karr, A. (2011). Deming, data and observational studies. *Significance*, 8(3):116–120.

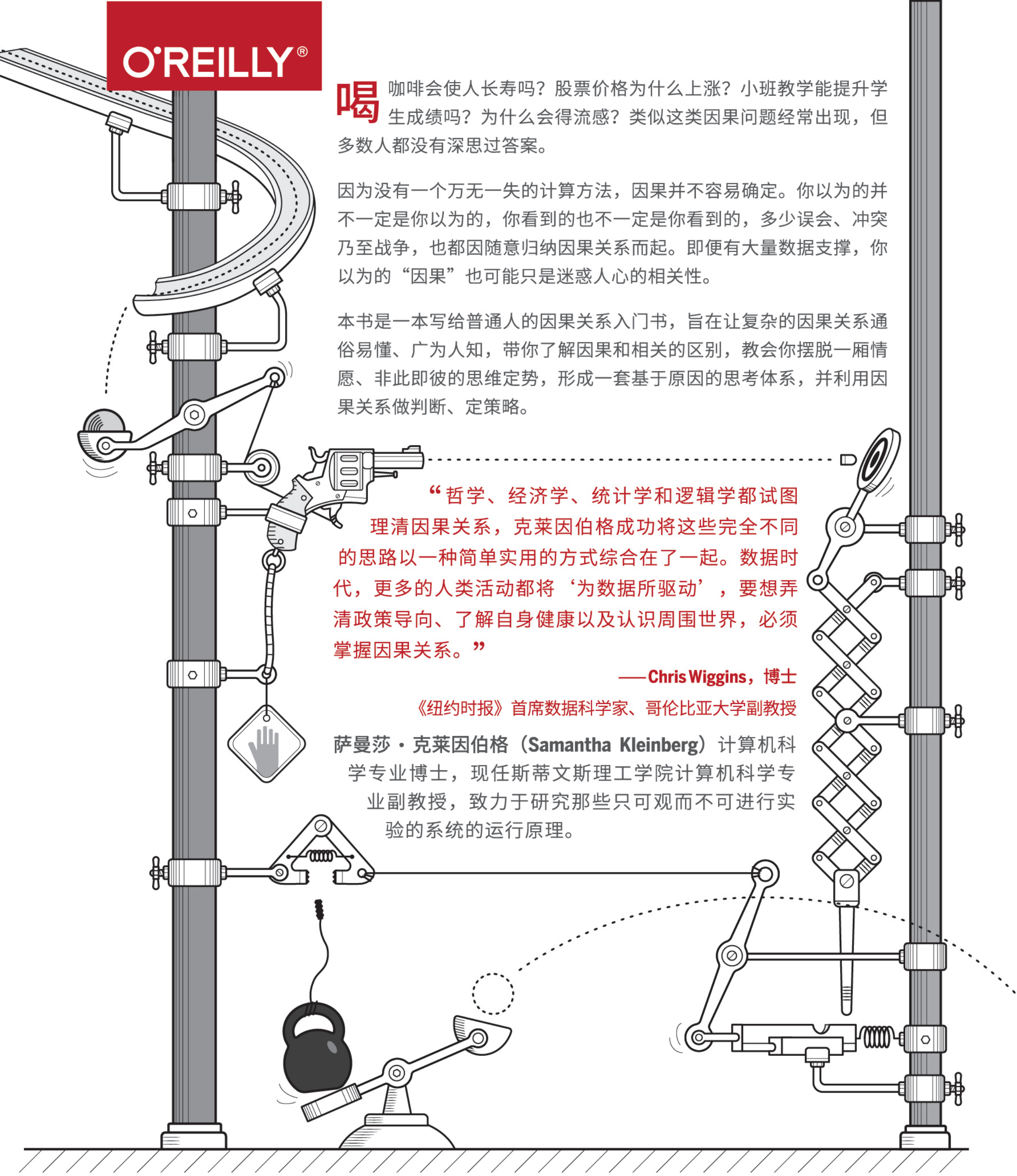
Yule, G. U. (1903). Notes on the Theory of Association of Attributes in Statistics. *Biometrika*, 2(2):121–134.

Zou, X., Tam, K.-P., Morris, M. W., Lee, S.-L., Lau, I. Y.-M., and Chiu, C.-Y. (2009). Culture as common sense: Perceived consensus versus personal beliefs as mechanisms of cultural influence. *Journal of Personality and Social Psychology*, 97(4):579–597.

关于作者

萨曼莎·克莱因伯格（Samantha Kleinberg）毕业于纽约大学，取得了计算机科学专业博士学位，现任斯蒂文斯理工学院计算机科学专业副教授，主要致力于开发一些方法，用来弄清那些只能观察而不能实验的系统的运行原理，其研究获得了美国国家科学基金会杰出青年学者奖（NSF CAREER Award）和杰姆斯·S. 麦克唐奈基金会复杂系统学者奖（James S. McDonnell Foundation Complex Systems Scholar Award）奖金资助。另著有 *Causality, Probability, and Time*。

O'REILLY®



喝咖啡会使人长寿吗？股票价格为什么上涨？小班教学能提升学生成绩吗？为什么会得流感？类似这类因果问题经常出现，但多数人都没有深思过答案。

因为没有一个万无一失的计算方法，因果并不容易确定。你以为的并不一定是你以为的，你看到的也不一定是你看到的，多少误会、冲突乃至战争，也都因随意归纳因果关系而起。即便有大量数据支撑，你以为的“因果”也可能只是迷惑人心的相关性。

本书是一本写给普通人的因果关系入门书，旨在让复杂的因果关系通俗易懂、广为人知，带你了解因果和相关的区别，教会你摆脱一厢情愿、非此即彼的思维定势，形成一套基于原因的思考体系，并利用因果关系做判断、定策略。

“哲学、经济学、统计学和逻辑学都试图理清因果关系，克莱因伯格成功将这些完全不同的思路以一种简单实用的方式综合在了一起。数据时代，更多的人类活动都将‘为数据所驱动’，要想弄清政策导向、了解自身健康以及认识周围世界，必须掌握因果关系。”

——**Chris Wiggins，博士**

《纽约时报》首席数据科学家、哥伦比亚大学副教授

萨曼莎·克莱因伯格（Samantha Kleinberg）计算机科学专业博士，现任斯蒂文斯理工学院计算机科学专业副教授，致力于研究那些只可观而不可进行实验的系统的运行原理。

分类建议　哲学／逻辑学

人民邮电出版社网址：www.ptpress.com.cn

O'Reilly Media, Inc.授权人民邮电出版社出版

ISBN 978-7-115-48518-2

9 787115 485182 >

ISBN 978-7-115-48518-2

定价：69.00元

看完了